Análisis de discurso en medios de comunicación digitales sobre corrupción en salud en Colombia (2022-2023), mediante técnicas de Procesamiento de Lenguaje Natural (PLN)

Michel Stivens Larrota Villalba

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá D.C., Colombia

20/05/2026

Análisis de discurso en medios de comunicación digitales sobre corrupción en salud en Colombia (2022-2023), mediante técnicas de Procesamiento de Lenguaje Natural (PLN)

Michel Stivens Larrota Villalba

Trabajo de grado presentado como requisito para optar al título de:

Magister en Ciencia de Datos

Directora:

Estefanía Mendoza Rodríguez

Modalidad:

Monografía

Universidad EAN

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá D.C., Colombia

20/05/2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá D.C., 20/05/2026

Dedicatoria

A mi hija y a mi madre, por su amor incondicional y por ser la fortaleza que me sostuvo durante todo este proceso.

A mi hermano, compañero de vida, cuyo ejemplo de perseverancia me recuerda que cada meta se construye paso a paso.

A mis abuelos, cuyo legado de honestidad y trabajo incansable ha orientado siempre mi comprensión del esfuerzo y la integridad como pilares del verdadero conocimiento.

"*El avance del conocimiento no pertenece a quienes se conforman con los caminos trazados, sino a quienes se atreven a crearlos.*" - Adaptación personal inspirada en el lema "*Inveniam viam aut faciam*".

Agradecimientos

En el transcurso de este proceso investigativo he tenido la oportunidad de contar con el apoyo de muchas personas que han contribuido a la culminación de esta tesis de maestría.

En primer lugar, quiero expresar mi especial y profundo agradecimiento a la profesora Estefanía Mendoza Rodríguez, directora del presente trabajo de grado, por su valioso acompañamiento académico y humano durante todo el proceso investigativo. Su orientación, disposición permanente y confianza fueron determinantes para superar las dificultades propias del desarrollo del proyecto y para consolidar una propuesta sólida y coherente. Más allá de su papel como directora, su apoyo personal y su compromiso con mi formación académica constituyeron un referente fundamental a lo largo de este trabajo.

Asimismo, agradezco a los docentes de la Universidad EAN que han contribuido a mi formación académica, incluyendo las dos especializaciones previamente cursadas y la presente Maestría en Ciencia de Datos. Su dedicación, exigencia académica y vocación formativa han permitido fortalecer las competencias profesionales y de investigación que hicieron posible la culminación de este trabajo.

Finalmente, reconozco a la Universidad EAN como una institución comprometida con el emprendimiento, la innovación y el pensamiento crítico, cuyo entorno académico ha facilitado un proceso continuo de aprendizaje que se proyecta hacia futuros desafíos académicos y profesionales.

Resumen

La presente investigación analiza el discurso mediático digital sobre la corrupción en el sector salud en Colombia durante el periodo 2022–2023 mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Para ello, se construyó un corpus de 518 noticias provenientes de seis medios digitales, seleccionados bajo criterios de relevancia investigativa y viabilidad técnica.

El estudio se desarrolló bajo el marco CRISP-DM, integrando modelado temático, análisis de sentimiento, reconocimiento de entidades y operacionalización de marcos narrativos. El modelado temático se implementó mediante Latent Dirichlet Allocation (LDA), complementado con enfoques alternativos para evaluar la estabilidad y consistencia de la estructura temática. Asimismo, el análisis de sentimiento se realizó mediante un modelo contextual basado en transformers, validado a través de recursos léxicos de referencia.

Los resultados evidencian la existencia de estructuras temáticas estables, patrones diferenciados de negatividad discursiva entre medios y configuraciones narrativas consistentes en la representación de la corrupción en salud. En conjunto, los hallazgos muestran que el discurso mediático presenta regularidades sistemáticas en la construcción de actores, responsabilidades y marcos interpretativos.

La investigación demuestra la viabilidad de un enfoque computacional triangulado para el análisis del discurso público a gran escala y aporta evidencia empírica sobre la construcción mediática de la corrupción en el sector salud de Colombia.

Palabras clave: Corrupción; sector salud; medios digitales; análisis del discurso; modelado temático; procesamiento de lenguaje natural; Colombia.

Abstract

This study analyzes digital media discourse on corruption in Colombia's healthcare sector during 2022–2023 using Natural Language Processing (NLP) techniques. To this end, a corpus of 518 news articles was compiled from six digital media outlets, selected based on criteria of research relevance and technical feasibility.

The study was conducted within the CRISP-DM framework, integrating topic modeling, sentiment analysis, named entity recognition, and the operationalization of narrative frames. Topic modeling was implemented using Latent Dirichlet Allocation (LDA), supplemented with alternative approaches to assess the stability and consistency of the thematic structure. Likewise, sentiment analysis was performed using a contextual transformer-based model, validated through reference lexical resources.

The results demonstrate the existence of stable thematic structures, distinct patterns of discursive negativity across media outlets, and consistent narrative configurations in the representation of corruption in the healthcare sector. Taken together, the findings show that media discourse exhibits systematic regularities in the construction of actors, responsibilities, and interpretive frames.

This research demonstrates the feasibility of a triangulated computational approach for large-scale public discourse analysis and provides empirical evidence on the media construction of corruption in Colombia's healthcare sector.

Keywords: Corruption; healthcare sector; digital media; discourse analysis; topic modeling; natural language processing; Colombia.

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Tabla de Siglas

Sigla	Significado
ACD	Análisis Crítico del Discurso
ADCSS	Análisis de Discurso sobre Corrupción en el Sector Salud
API	Application Programming Interface
ARI	Adjusted Rand Index
BERT	Bidirectional Encoder Representations from Transformers
BL	Backlinks (enlaces entrantes a un sitio web)
CRISP-DM	Cross-Industry Standard Process for Data Mining (Proceso estándar para minería de datos)
C_npmi	Normalized Pointwise Mutual Information
C_v	Métrica de coherencia temática basada en coocurrencia mediante ventana deslizante
DA	Domain Authority
DANE	Departamento Administrativo Nacional de Estadística
EAN	Universidad EAN
EPS	Entidad Promotora de Salud
GPT	Generative Pre-trained Transformer
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HDP	Hierarchical Dirichlet Process
HTTP	HyperText Transfer Protocol
IA	Inteligencia Artificial
JSON	JavaScript Object Notation
KW	Keywords
LDA	Latent Dirichlet Allocation
ML	Machine Learning
NER	Named-Entity Recognition (Reconocimiento de Entidades Nombradas)
NLP	Natural Language Processing (equivalente en inglés de PLN)
NLTK	Natural Language Toolkit
NMI	Normalized Mutual Information
NRC	NRC Emotion Lexicon
PLN	Procesamiento de Lenguaje Natural
RAM	Random Access Memory
REST	Representational State Transfer
RoBERTa	Robustly Optimized BERT Pretraining Approach
RSS	Really Simple Syndication
SGSSS	Sistema General de Seguridad Social en Salud
SSD	Solid-State Drive
TD	Topic Diversity (Diversidad temática)
TF-IDF	Term Frequency–Inverse Document Frequency
TRA	Tráfico estimado (Traffic)
UMAP	Uniform Manifold Approximation and Projection

Nota. Siglas utilizadas en el documento.

1. Introducción

La corrupción en el sector salud constituye una problemática estructural que afecta la eficiencia institucional, debilita la confianza ciudadana y limita el acceso equitativo a servicios esenciales (Transparencia por Colombia, 2022; Vian, 2019). En el contexto colombiano, diversos estudios e informes institucionales han documentado la existencia de redes de intermediación, fraudes contractuales y malversación de recursos que impactan la gestión del Sistema General de Seguridad Social en Salud (Contraloría General de la República, 2023; Guerrero et al., 2019). Esta situación ha generado un creciente interés por comprender no solo las dimensiones administrativas y jurídicas del fenómeno, sino también su representación en el discurso público.

En este escenario, los medios de comunicación digitales desempeñan un papel central en la configuración de la opinión pública, ya que no se limitan a informar sobre los hechos, sino que contribuyen a su encuadre interpretativo, influyendo en las percepciones sociales, políticas y éticas de la corrupción (Boydstun & Shafer, 2017; Rodríguez & García, 2021). El periodo comprendido entre 2022 y 2023 se caracteriza por una alta visibilización mediática de presuntas irregularidades en el sector salud, así como por el impacto social asociado al cambio de gobierno y al debate público en torno a propuestas de reforma estructural del sistema, lo que intensificó la producción y circulación de contenidos periodísticos sobre el tema (El Espectador, 2023b; Ministerio de Salud y Protección Social, 2023). En esta cobertura, no solo resulta relevante qué hechos se informan, sino también qué actores son visibilizados como protagonistas del relato público, entre ellos actores políticos, Entidades Promotoras de Salud (EPS), pacientes y organismos de control.

Si bien existe abundante información institucional y financiera sobre los casos de corrupción en salud, la evidencia empírica acerca de las narrativas mediáticas que los acompañan resulta limitada. La literatura reconoce que el discurso periodístico no es neutral y

que responde a lógicas editoriales, contextos de polarización y dinámicas de encuadre ideológico (Greussing & Boomgaarden, 2016; Pérez, 2023). En este sentido, el análisis sistemático del discurso mediático permite identificar patrones recurrentes en la forma como se tematiza y valora la corrupción en salud dentro del espacio público digital.

En respuesta a esta brecha, la presente investigación tiene como propósito identificar patrones lingüísticos, temáticos y emocionales en el tratamiento noticioso de la corrupción en el sector salud en Colombia mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Para ello, se analizó un corpus de noticias digitales publicadas entre 2022 y 2023 en medios nacionales y regionales, seleccionados con base en criterios de relevancia investigativa, alcance digital y viabilidad técnica de extracción. A partir de este corpus, el estudio busca caracterizar los principales ejes discursivos presentes en la cobertura mediática, examinar la carga emocional asociada a dichos ejes y analizar las dinámicas de visibilización de actores, contribuyendo a una comprensión empírica de la representación mediática de este fenómeno.

La pregunta que orienta esta investigación es la siguiente: ¿Qué patrones temáticos, emocionales y de visibilización de actores pueden identificarse en el discurso mediático digital sobre la corrupción en el sector salud en Colombia (2022–2023) mediante técnicas de Procesamiento de Lenguaje Natural (PLN)?

El documento se estructura de la siguiente manera: en primer lugar, se presentan los objetivos y la justificación de la investigación; posteriormente, se desarrolla el marco teórico, abordando los conceptos de corrupción en el sector salud, medios digitales, análisis del discurso y PLN. A continuación, se describe la metodología empleada, incluyendo el diseño analítico y los procedimientos de modelado y evaluación. Seguidamente, se exponen los resultados y su discusión, y finalmente se presentan las conclusiones y proyecciones para futuras investigaciones.

2. Objetivos

2.1. Objetivo general

Analizar el discurso mediático sobre la corrupción en el sector salud en Colombia (2022–2023) publicado en medios digitales nacionales y regionales, mediante técnicas de Procesamiento de Lenguaje Natural (PLN), con el fin de identificar patrones temáticos, variaciones en el tono emocional, marcos narrativos y dinámicas de visibilización de actores.

2.2. Objetivos específicos

- Caracterizar el marco discursivo asociado a la corrupción en el sector salud en Colombia, integrando enfoques de teoría del discurso y framing mediático, con el fin de establecer las categorías conceptuales que orientan el análisis mediante técnicas de Procesamiento de Lenguaje Natural.
- Identificar y modelar la estructura temática latente del corpus mediante Latent Dirichlet Allocation (LDA), validando la selección óptima del número de tópicos a través de métricas de coherencia semántica y diversidad temática, y contrastando los resultados con modelos alternativos para evaluar su estabilidad y plausibilidad estructural.
- Analizar las variaciones en el tono emocional del discurso y las dinámicas de visibilización de actores en la cobertura mediática mediante modelos de análisis de sentimiento y técnicas de reconocimiento de entidades, evaluando diferencias por medio, tópico y periodo temporal.
- Integrar los hallazgos temáticos, emocionales y de atribución de responsabilidad para operacionalizar marcos narrativos como aproximación computacional al framing mediático, evaluando su consistencia y estabilidad mediante análisis comparativos y validación estadística.

3. Justificación

La corrupción en el sector salud en Colombia afecta la calidad, accesibilidad y equidad de la atención, con impactos desproporcionados sobre poblaciones vulnerables (Guerrero et al., 2019; Vian, 2019). Aunque existen informes institucionales y de control fiscal sobre casos, montos y procesos (Contraloría General de la República, 2023; Transparencia por Colombia, 2022), persiste un vacío en el análisis sistemático de las narrativas mediáticas digitales que enmarcan estos hechos durante el periodo 2022–2023.

El estudio se justifica porque los medios digitales no solo informan, sino que encuadran y jerarquizan la información, influyendo en percepciones, emociones y atribuciones de responsabilidad en la opinión pública (Boydstun & Shafer, 2017; Rodríguez & García, 2021). Analizar estos encuadres mediante métodos reproducibles permite identificar sesgos, marcos interpretativos y patrones de visibilización de actores, aportando evidencia útil para fortalecer la transparencia y la rendición de cuentas.

Desde el punto de vista metodológico, la aplicación de técnicas de Procesamiento de Lenguaje Natural a noticias digitales permite analizar grandes volúmenes de información no estructurada para detectar estructuras temáticas, polaridad discursiva y dinámicas temporales (Blei, 2012; Medhat et al., 2014; Yin et al., 2021). Este enfoque permite transformar corpus periodísticos extensos en evidencia analítica sistemática, favoreciendo la extracción de conocimiento a partir de datos textuales que, por su volumen y heterogeneidad, serían difíciles de abordar únicamente mediante revisión manual. Asimismo, aunque el PLN no elimina por completo los sesgos del análisis, sí contribuye a reducir la dependencia de apreciaciones individuales del investigador al aplicar criterios de procesamiento, clasificación y comparación de manera homogénea, trazable y reproducible sobre la totalidad del corpus.

La literatura reciente respalda el uso de enfoques computacionales no supervisados para examinar narrativas sobre corrupción en entornos digitales (Li et al., 2020) y muestra que

las tecnologías emergentes pueden tanto fortalecer como limitar los procesos anticorrupción, por lo que su evaluación requiere evidencia empírica rigurosa y contextualizada (Adam & Fazekas, 2021).

En términos aplicados, los resultados son relevantes para periodistas, entes de control y tomadores de decisión, al proporcionar métricas comparables sobre tono, marcos narrativos y visibilidad de actores. Asimismo, complementan los mecanismos tradicionales de vigilancia institucional con evidencia discursiva que puede contribuir a anticipar dinámicas de opinión pública y riesgos de desinformación (Superintendencia Nacional de Salud, 2024). Para la academia, el estudio aporta a un campo aún incipiente en la región del análisis computacional del discurso mediático sobre corrupción, mediante un protocolo replicable y validable.

Finalmente, la pertinencia temporal se relaciona con el debate público sobre la reforma al sistema de salud intensificado desde 2023, que incrementó la cobertura mediática y la polarización del tema (Ministerio de Salud y Protección Social, 2023). Analizar las narrativas construidas en ese periodo proporciona una línea base empírica para el seguimiento ciudadano y la evaluación de políticas orientadas a fortalecer la transparencia y la confianza institucional.

4. Marco Teórico

Este capítulo establece los fundamentos conceptuales que enmarcan el análisis del discurso mediático digital sobre la corrupción en el sector salud en Colombia durante el periodo 2022–2023, orientando las decisiones analíticas adoptadas y garantizando su coherencia con un corpus periodístico heterogéneo en español. El marco integra enfoques del análisis del discurso y del framing mediático con técnicas de Procesamiento de Lenguaje Natural (PLN) aplicadas al estudio de noticias digitales, asegurando correspondencia entre conceptos teóricos, variables analíticas y criterios de operacionalización.

Asimismo, delimita los supuestos interpretativos y criterios de validez que sustentan el análisis, articulando el fenómeno sustantivo con los marcos conceptuales y las decisiones analíticas empleadas. Esta integración permite alinear los fundamentos teóricos con los objetivos de investigación y los estándares de calidad metodológica adoptados.

4.1. Propósito del marco teórico

El marco parte del reconocimiento de que la corrupción en el sector salud constituye un fenómeno estructural, multicausal y discursivamente complejo, cuya visibilidad y legitimación dependen en gran medida del tratamiento que recibe en los medios de comunicación digitales. En este contexto, la prensa digital no se limita a transmitir hechos, sino que selecciona, jerarquiza y encuadra los acontecimientos, contribuyendo a la configuración de climas de opinión y percepciones ciudadanas sobre la integridad institucional (Boydstun & Shafer, 2017; Rodríguez & García, 2021).

La investigación se sustenta en la teoría del discurso y en el framing mediático como marcos conceptuales que permiten analizar cómo el lenguaje configura significados, define actores y delimita responsabilidades en el debate público (Entman, 1993; Fairclough, 2013). Estos enfoques se complementan con la perspectiva computacional del Procesamiento de Lenguaje Natural (PLN), que posibilita el análisis sistemático de grandes corpus de noticias

mediante algoritmos orientados a identificar patrones temáticos y variaciones en el tono emocional del discurso (Camacho-Collados & Pilehvar, 2018; Yin et al., 2021).

La estructura del marco se organiza en torno a tres ejes principales:

1. **El fenómeno social:** la corrupción en el sector salud y su relevancia pública reciente, sustentado en estudios sobre corrupción sistémica, gobernanza y riesgos institucionales (Guerrero et al., 2019; Rose-Ackerman & Palifka, 2016; Vian, 2019).
2. **El componente comunicativo:** el papel de los medios digitales en la construcción del discurso público, a partir del análisis crítico del discurso y la teoría del framing, que explican cómo los medios seleccionan, jerarquizan y encuadran los hechos noticiosos (Boydstun & Shafer, 2017; de Vreese, 2019; Entman, 1993; Fairclough, 2013; van Dijk, 2015).
3. **El componente técnico-analítico:** el uso de técnicas de Procesamiento de Lenguaje Natural para el análisis empírico del discurso mediático, particularmente el modelado temático no supervisado mediante Latent Dirichlet Allocation (LDA) (Blei et al., 2003) y el análisis de sentimiento basado en modelos de lenguaje contextualizados de tipo transformer, aplicados a la estimación de la polaridad y orientación emocional del discurso (Devlin et al., 2019; Yin et al., 2021). Este eje incorpora además métricas de coherencia temática y diversidad tópica como criterios conceptuales de validación del modelado (Dieng et al., 2020; Röder et al., 2015).

En conjunto, estos componentes articulan fenómeno, teoría y técnica analítica, proporcionando una base conceptual consistente para la interpretación de los resultados. Esta organización permite transitar desde la comprensión del objeto de estudio hasta los criterios analíticos que sustentan su operacionalización empírica.

4.2. Corrupción en salud en Colombia

La corrupción en el sector salud constituye una de las manifestaciones más relevantes del debilitamiento institucional y de la erosión de la confianza pública en Colombia. La literatura especializada ha documentado su expresión en prácticas como desviación de recursos, captura de rentas y manipulación de procesos contractuales, las cuales comprometen directamente la prestación de servicios esenciales y el acceso efectivo a la atención sanitaria (Contraloría General de la República, 2023; Transparencia por Colombia, 2022). Desde el ámbito académico, este fenómeno ha sido caracterizado como una forma de corrupción sistémica en la que conductas ilegales o éticamente cuestionables se integran en rutinas organizacionales y políticas, generando dinámicas persistentes de impunidad (Rose-Ackerman & Palifka, 2016; Vian, 2019).

Desde una perspectiva funcional, la corrupción en salud puede entenderse como el uso indebido del poder público o privado con funciones delegadas para beneficio particular dentro de la cadena de provisión, contratación o control de los servicios sanitarios. En Colombia, esta dinámica se ve intensificada por la estructura del Sistema General de Seguridad Social en Salud, que articula actores públicos, privados y mixtos bajo esquemas de intermediación contractual complejos. Guerrero et al. (2019) señalan que la fragmentación institucional y la debilidad de los mecanismos de supervisión han favorecido prácticas como sobrecostos, falsificación de facturas y sobornos en los procesos de adjudicación contractual.

4.2.1. Definiciones y tipologías

Vian (2019) clasifica la corrupción en salud en seis categorías principales: sobornos, desvío de fondos, nepotismo, fraude en licitaciones, comercialización ilegal de insumos o medicamentos y falsificación de datos clínicos o financieros. Transparency International (2021) amplía este marco al incluir prácticas como clientelismo político en la asignación de recursos, abuso de autoridad administrativa y colusiones empresariales que distorsionan la competencia.

Estas tipologías se utilizan en el presente estudio como marco conceptual de referencia para interpretar los patrones temáticos y narrativos identificados en el discurso mediático, desde denuncias puntuales hasta representaciones de redes estructuradas de corrupción.

4.2.2. Relevancia pública 2022–2023

Durante el periodo 2022–2023, la corrupción en el sector salud alcanzó una elevada visibilidad mediática debido a la convergencia de varios factores: el debate nacional sobre la reforma del sistema de salud, la continuidad de investigaciones relacionadas con irregularidades contractuales asociadas a la pandemia y su poscrisis, y la exposición de casos regionales que involucraron tanto a entidades públicas como privadas. Según la Contraloría General de la República (2023), en este periodo se abrieron más de 200 investigaciones por presunto desvío de recursos y contratación irregular.

Diversos medios nacionales abordaron el tema como asunto de alto interés público, frecuentemente vinculado a controversias políticas y debates sobre eficiencia estatal (El Espectador, 2023b, 2023a; La República, 2023). Informes de Transparencia por Colombia (2022) reportan un incremento significativo de menciones mediáticas a casos de corrupción sanitaria entre 2021 y 2023, con énfasis en contratación, sobrecostos y fallas de vigilancia.

Este contexto intensamente mediatizado convierte la corrupción en salud en un objeto discursivo susceptible de análisis empírico, más allá de su dimensión administrativa o jurídica.

4.2.3. Corrupción, salud y confianza institucional

La relación entre corrupción y salud pública trasciende los efectos financieros y se proyecta sobre dimensiones sociales y simbólicas del sistema. La exposición reiterada a escándalos de corrupción contribuye a la disminución de la confianza ciudadana en las instituciones sanitarias y a la erosión de la percepción de legitimidad del sistema (Gaitán et al., 2020; Jain et al., 2022). Este deterioro se asocia con menores niveles de respaldo a políticas públicas y con cambios en la disposición ciudadana a interactuar con el sistema de salud.

En este contexto, la dimensión discursiva adquiere especial relevancia. Los medios actúan como intermediarios entre los hechos y su interpretación social, influyendo en la asignación simbólica de responsabilidades y en la percepción de impunidad o rendición de cuentas. El análisis del discurso digital, apoyado en enfoques lingüísticos y computacionales, permite aportar evidencia empírica sobre estos procesos de construcción simbólica del poder, la responsabilidad pública y la confianza institucional (Adam & Fazekas, 2021; Fairclough, 2013).

En síntesis, la corrupción en salud se conceptualiza como un fenómeno estructural de alto impacto social y como un objeto discursivo susceptible de análisis sistemático mediante técnicas de PLN, proporcionando el contexto necesario para vincular su representación mediática con la estrategia analítica del estudio.

4.3. Medios digitales y opinión pública

Los medios digitales se han consolidado como actores centrales en la configuración de la opinión pública contemporánea. En Colombia, su influencia trasciende la difusión de información y se proyecta como un proceso activo de construcción de significados sobre los asuntos públicos (Moyano & Salazar, 2022). Este fenómeno se enmarca en la transformación del sistema comunicativo hacia un entorno mediático en red, caracterizado por la interconexión, la inmediatez y la fragmentación de audiencias (Castells, 2009; Couldry & Hepp, 2017).

Desde la perspectiva de la agenda setting, los medios no solo informan, sino que determinan qué temas adquieren centralidad en el debate público (McCombs & Shaw, 1972). Esta función se complementa con los efectos de encuadre o framing, entendidos como los mecanismos mediante los cuales los medios seleccionan ciertos aspectos de la realidad y los presentan bajo interpretaciones específicas (de Vreese, 2019; Entman, 1993). En contextos de corrupción, estos encuadres influyen en la percepción social del fenómeno, ya sea como

desviación individual, falla administrativa o problema estructural de gobernanza (Boydstun & Shafer, 2017).

En el caso colombiano, la cobertura digital de la corrupción ha mostrado tendencias hacia la atribución de responsabilidad a actores políticos y hacia narrativas de indignación pública, dinámicas que pueden intensificar procesos de polarización y afectar la confianza institucional (Arroyave & Barrios, 2023; López-Londoño & Molinares, 2021). Esta dimensión resulta especialmente relevante en el sector salud, donde los marcos mediáticos influyen en la percepción de legitimidad del sistema y en la evaluación ciudadana de la gestión pública.

Desde esta perspectiva, el análisis sistemático del discurso mediático digital permite examinar empíricamente cómo se estructuran las narrativas, qué actores son visibilizados y cómo se configura la carga emocional del debate público. El uso de técnicas de Procesamiento de Lenguaje Natural permite abordar estos procesos a gran escala con criterios de rigor y reproducibilidad.

En coherencia con este enfoque, la investigación adopta el framing mediático como marco interpretativo central para comprender cómo los medios organizan y presentan la información sobre la corrupción en salud y cómo estas representaciones contribuyen a la construcción social del fenómeno.

4.4. Teoría del discurso y framing

El análisis del discurso mediático requiere comprender no solo qué temas abordan los medios, sino también cómo estos son interpretados y presentados al público. La teoría del framing sostiene que los medios seleccionan ciertos aspectos de la realidad y los hacen más prominentes dentro del discurso informativo, promoviendo interpretaciones específicas sobre los acontecimientos (Entman, 1993).

Los marcos interpretativos actúan como estructuras que organizan el significado de los hechos y orientan la comprensión pública de los problemas sociales (Goffman, 1974;

Scheufele, 1999). En el ámbito periodístico, estos marcos se manifiestan mediante narrativas recurrentes, énfasis temáticos y atribuciones de responsabilidad que estructuran la representación mediática de los fenómenos.

La combinación del Análisis Crítico del Discurso (ACD) y la teoría del framing permite abordar el discurso desde una perspectiva multinivel. El ACD enfatiza la dimensión lingüística, ideológica y sociopolítica del texto, entendiendo el lenguaje como práctica social situada (Fairclough, 2013; van Dijk, 2015), mientras que el framing examina las estructuras interpretativas que orientan la comprensión pública de los acontecimientos (de Vreese, 2019; Entman, 1993). Esta articulación resulta especialmente pertinente para el estudio de la corrupción en salud, donde la representación mediática no solo describe hechos, sino que contribuye a definir responsabilidades, evaluar moralmente a los actores e interpretar el fenómeno como desviación individual o problema estructural (Arroyave & Barrios, 2023; Boydstun & Shafer, 2017).

En investigaciones recientes, el análisis computacional del discurso se ha utilizado para operacionalizar categorías teóricas del framing en grandes corpus textuales, permitiendo identificar patrones léxicos, temáticos y emocionales que reflejan regularidades discursivas en la cobertura mediática (Camacho-Collados & Pilehvar, 2018; Yin et al., 2021). No obstante, estos modelos no detectan frames en sentido hermenéutico estricto, sino regularidades estadísticas que deben interpretarse a la luz de marcos teóricos previamente definidos (Dieng et al., 2020; Röder et al., 2015).

En este estudio, los encuadres se conceptualizan como patrones recurrentes en la forma en que las noticias presentan el problema público, atribuyen responsabilidades y evalúan los hechos. Dado que no se realiza etiquetado manual de frames, estos se aproximan indirectamente mediante tres dimensiones analíticas coherentes con la literatura: (i) los tópicos dominantes identificados por modelado temático no supervisado (Blei et al., 2003), (ii) la polaridad y variación del tono emocional entre medios y periodos, estimada mediante modelos

de lenguaje contextualizados (Devlin et al., 2019; Yin et al., 2021), y (iii) la visibilidad y recurrencia de actores asociados a responsabilidad o conflicto, en línea con enfoques de análisis de actores discursivos (van Dijk, 2015).

Estas dimensiones complementan, pero no sustituyen, el análisis interpretativo propio del ACD. Los patrones identificados se interpretan a partir de categorías teóricas previamente delimitadas, tales como corrupción como desviación individual, falla institucional o crisis estructural de gobernanza.

De este modo, el enfoque de discurso y framing orienta la interpretación sustantiva de los resultados y permite transformar el corpus periodístico en variables analíticas cuantificables, articulando teoría crítica del discurso con métodos computacionales contemporáneos.

4.4.1. Tipología de marcos narrativos

En el análisis del discurso mediático, los marcos interpretativos constituyen estructuras narrativas que organizan la manera en que los medios presentan los acontecimientos, definen los problemas públicos, atribuyen responsabilidades y sugieren evaluaciones morales o políticas (de Vreese, 2019; Entman, 1993). En el periodismo político y de investigación, estos encuadres permiten comprender cómo se representan fenómenos complejos como la corrupción.

La literatura identifica tipologías recurrentes de framing, entre ellas marcos asociados al conflicto político, las consecuencias económicas, la atribución de responsabilidad legal, el impacto humano y las evaluaciones morales (Boydstun & Shafer, 2017; de Vreese, 2019). Estas categorías funcionan como esquemas interpretativos que simplifican la complejidad de los fenómenos sociales y facilitan su comprensión pública.

Con base en estos aportes, la presente investigación adopta una tipología analítica de cinco marcos narrativos que orientan la interpretación de los resultados obtenidos mediante técnicas de PLN. Estos marcos no se identifican mediante etiquetado manual directo, sino que

se aproximan indirectamente a partir de la convergencia entre patrones temáticos, variaciones en el tono emocional y visibilidad de actores dentro del corpus analizado.

Tabla 1. Tipología de marcos narrativos utilizados en el análisis

Marco narrativo	Definición conceptual	Indicadores discursivos aproximados
Conflicto político	Presenta la corrupción como resultado de disputas entre actores políticos o institucionales.	Confrontación entre actores, acusaciones cruzadas, disputas institucionales.
Consecuencia económica	Enfatiza los efectos financieros o administrativos derivados de actos de corrupción.	Pérdidas económicas, recursos desviados, impacto presupuestal.
Judicial / legal	Interpreta los hechos desde la perspectiva de investigaciones, sanciones o procesos judiciales.	Fiscalía, tribunales, investigaciones, imputaciones.
Impacto humano	Destaca las consecuencias de la corrupción sobre ciudadanos, pacientes o usuarios del sistema de salud.	Afectación a pacientes, servicios médicos, acceso a tratamientos.
Moral / ético	Presenta la corrupción como una transgresión normativa o moral que genera indignación pública.	Condena moral, indignación, juicios éticos sobre actores.

Nota. Elaboración propia basada en la literatura sobre framing mediático (Boydstun & Shafer, 2017; de Vreese, 2019). Las categorías y definiciones constituyen una síntesis analítica adaptada al contexto del estudio, y los indicadores discursivos corresponden a operacionalizaciones aproximadas diseñadas para su identificación indirecta mediante técnicas de PLN.

Finalmente, estos marcos no constituyen categorías mutuamente excluyentes ni etiquetas rígidas aplicadas a cada documento. Funcionan como una aproximación interpretativa destinada a sintetizar patrones discursivos identificados mediante análisis computacional, articulando la evidencia empírica con las categorías conceptuales del análisis del discurso y del framing mediático.

4.5. Procesamiento de Lenguaje Natural (PLN) para análisis de discurso

El Procesamiento de Lenguaje Natural (PLN) es un campo interdisciplinario que integra lingüística computacional, estadística e inteligencia artificial con el propósito de modelar y analizar el lenguaje humano a gran escala. En las ciencias sociales, su aplicación permite

transformar grandes volúmenes de texto en estructuras analíticas observables, facilitando la identificación sistemática de patrones semánticos, estructuras temáticas y regularidades discursivas (Camacho-Collados & Pilehvar, 2018; Eisenstein, 2019).

En el estudio del discurso mediático, el PLN posibilita superar las limitaciones del análisis manual mediante aproximaciones empíricas escalables, en las que algoritmos supervisados y no supervisados permiten detectar tópicos dominantes, estimar polaridad emocional e identificar recurrencias léxicas asociadas a actores y eventos. Esta capacidad resulta especialmente relevante en investigaciones sobre corrupción debido al volumen y dinamismo de las publicaciones digitales.

El desarrollo reciente de modelos de lenguaje basados en arquitecturas transformer (Vaswani et al., 2017), como BERT y sus variantes multilingües (Devlin et al., 2019), ha mejorado la estimación contextual del significado al capturar dependencias sintácticas y relaciones semánticas complejas. Estas representaciones han ampliado el uso del PLN en análisis de discurso político, comunicación digital y estudios de gobernanza (Lindgren, 2022; Zhang et al., 2023).

De manera complementaria, el modelado temático no supervisado, particularmente mediante Latent Dirichlet Allocation (LDA) (Blei et al., 2003), permite identificar estructuras latentes en corpus extensos. En años recientes, se han desarrollado enfoques híbridos que combinan modelado temático tradicional con embeddings contextuales derivados de modelos transformer, como BERTopic, que integra representaciones semánticas con técnicas de reducción de dimensionalidad y clustering (Grootendorst, 2022). Estas estrategias permiten contrastar estructuras temáticas desde distintas aproximaciones algorítmicas.

En el presente estudio, el PLN se emplea con un enfoque integrador que articula modelado temático no supervisado, estimación automatizada de polaridad emocional y análisis de recurrencia de actores discursivos. La combinación de estos componentes permite aproximar las dimensiones temáticas, emocionales y actorales del discurso mediático sobre

corrupción en salud, manteniendo coherencia con las categorías teóricas previamente delimitadas.

La elección de estas técnicas se fundamenta en tres criterios principales: (i) escalabilidad frente a un corpus periodístico extenso y heterogéneo en español, (ii) reproducibilidad mediante métricas de coherencia y validación intermodelo, y (iii) integración entre enfoques del análisis del discurso y ciencia de datos aplicada.

Dado que el desempeño de los modelos depende de la calidad del texto de entrada, el siguiente apartado describe los criterios de preparación y depuración del corpus periodístico, procesos esenciales para garantizar la validez analítica del modelado posterior.

4.6. Preparación del corpus y criterios de depuración

La calidad del modelado temático y del análisis de sentimiento depende directamente de la preparación del corpus. En estudios de PLN aplicados a noticias digitales, la limpieza del texto constituye una decisión metodológica que incide en la coherencia semántica de los tópicos, la estabilidad de los modelos no supervisados y la validez interpretativa de los resultados (Eisenstein, 2019).

En la presente investigación, la preparación de los datos se orientó a reducir ruido estructural, evitar redundancias y asegurar la pertinencia temática del corpus. Este proceso incluyó las siguientes etapas principales:

- Eliminación de duplicados mediante técnicas de hash y verificación probabilística, con el fin de evitar la sobreponderación de contenidos replicados.
- Limpieza de boilerplate, suprimiendo bloques editoriales repetitivos, menús de navegación, enlaces internos y elementos no informativos.
- Normalización básica del texto, incluyendo conversión a minúsculas y eliminación de caracteres no relevantes.

- Tokenización, entendida como la segmentación del texto en unidades léxicas mínimas necesarias para el análisis estadístico.
- Eliminación de stopwords para excluir términos funcionales de alta frecuencia sin contenido semántico sustantivo.
- Lematización, aplicada principalmente al modelado temático, con el fin de reducir la variación morfológica y mejorar la cohesión semántica.
- Construcción de bigramas y trigramas para capturar expresiones compuestas relevantes (por ejemplo, “corrupción_salud”, “reforma_sistema_salud”).
- Filtrado temático estricto basado en la intersección conceptual Salud ∩ Corrupción, garantizando la alineación con el objeto de estudio.
- Exclusión de documentos excesivamente breves o semánticamente vacíos tras la limpieza, evitando ruido residual.
- Control del rango temporal 2022–2023 como delimitación analítica del corpus.

Estas decisiones permitieron estructurar un corpus coherente y temáticamente consistente, adecuado para el modelado no supervisado y el análisis posterior. La preparación del texto se concibe, por tanto, como una etapa de estandarización semántica que condiciona directamente la calidad de los resultados obtenidos.

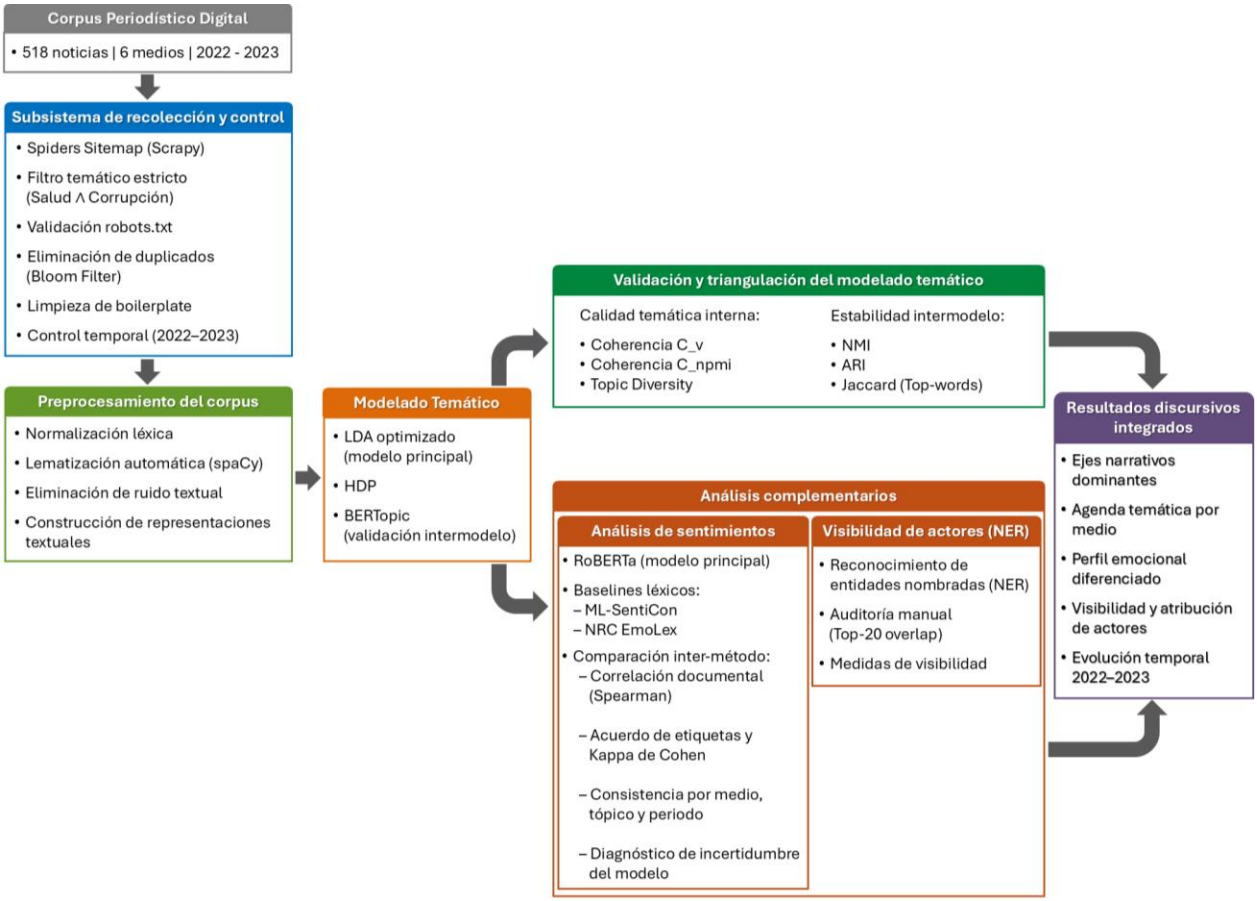
Adicionalmente, la fase de recolección y preclasificación temática incorporó listas de palabras clave adaptadas al contexto colombiano de corrupción en salud, incluyendo expresiones coloquiales, nombres de escándalos y combinaciones léxicas de uso frecuente en el discurso periodístico nacional, con el fin de reducir omisiones temáticas durante la captura y el filtrado del corpus.

4.7. Modelado temático y comparación

El modelado temático constituye una de las técnicas más utilizadas en PLN para identificar patrones latentes de significado en grandes volúmenes de texto. Permite agrupar

documentos según los temas que los componen, ofreciendo una representación estructurada del discurso mediático. En estudios sobre comunicación política y corrupción, resulta especialmente útil para detectar tópicos dominantes, comparar marcos narrativos y analizar su evolución temporal (Lindgren, 2022). En coherencia con el flujo metodológico sintetizado en la Figura 1, la investigación adopta un esquema comparativo que combina una línea base clásica con modelos alternativos para triangulación.

Figura 1. Arquitectura metodológica del análisis de discurso mediático digital (2022–2023)



Nota. El modelo LDA constituye el modelo principal validado estadísticamente de la investigación. Los modelos HDP y BERTopic se emplean como estrategias de triangulación para evaluar estabilidad temática bajo distintos supuestos metodológicos. La arquitectura integra modelado temático, análisis de sentimiento y visibilidad de actores.

El modelo Latent Dirichlet Allocation (LDA) (Blei et al., 2003) se adopta como línea base debido a su interpretabilidad, robustez y amplia utilización en estudios de minería de textos y análisis de discurso. Este enfoque representa los documentos como combinaciones probabilísticas de tópicos, aunque asume independencia contextual entre palabras, lo que puede limitar la coherencia semántica en textos breves o altamente figurativos (Wallach et al., 2009).

Como complemento, se consideran enfoques alternativos desarrollados a partir de representaciones contextuales del lenguaje. Entre ellos, BERTopic (Grootendorst, 2022) combina embeddings derivados de modelos transformer con técnicas de reducción de dimensionalidad y clustering basado en densidad, lo que permite identificar agrupaciones semánticamente coherentes sensibles al contexto lingüístico. Asimismo, el modelo HDP (Hierarchical Dirichlet Process) permite estimar de forma no paramétrica el número potencial de tópicos sin requerir su definición previa (Teh et al., 2006).

En esta investigación, LDA optimizado constituye el modelo principal de identificación temática, mientras que BERTopic y HDP se emplean como estrategias de contraste para evaluar estabilidad y coherencia temática. Esta triangulación permite comparar resultados bajo supuestos estadísticos distintos: probabilísticos en LDA, no paramétricos en HDP y semántico-contextuales en BERTopic.

La calidad del modelado se evaluó mediante métricas de coherencia temática y diversidad. En particular, se utilizó la coherencia C_v, diseñada para aproximar la interpretabilidad humana de los tópicos, donde valores más altos indican mayor cohesión semántica (Röder et al., 2015). De manera complementaria, se calculó la coherencia C_npmi como medida adicional de robustez léxica basada en asociaciones entre términos (Bouma, 2009; Lau et al., 2014), así como la diversidad temática (Topic Diversity), que cuantifica la proporción de palabras únicas entre los términos principales de los tópicos (Dieng et al., 2020).

Dado que el estudio no dispone de etiquetas manuales para validación externa, se incorporó una medida de concordancia entre particiones como parte de la triangulación entre modelos no supervisados. En particular, se comparó la asignación temática por documento entre LDA y BERTopic mediante NMI (Normalized Mutual Information) y ARI (Adjusted Rand Index), métricas que cuantifican el grado de acuerdo estructural entre agrupamientos (Hubert & Arabie, 1985; Strehl & Ghosh, 2003; Vinh et al., 2010). De forma complementaria, se evaluó la similitud léxica entre tópicos mediante el índice de Jaccard aplicado a los términos principales (top-words). Estas medidas permiten evaluar consistencia metodológica, aunque no exactitud absoluta.

Considerando que BERTopic utiliza HDBSCAN como algoritmo de agrupamiento, el cual puede clasificar algunos documentos como ruido, la concordancia se estimó únicamente sobre los textos con asignación temática válida y reportando explícitamente la proporción de documentos no asignados. Este procedimiento asegura trazabilidad analítica y evita sesgos derivados de observaciones excluidas.

La integración de métricas de coherencia, diversidad y concordancia responde a la ausencia de una referencia externa (ground truth), permitiendo evaluar la consistencia estructural del modelado temático. La arquitectura metodológica se implementa mediante un pipeline reproducible del proyecto ADCSS, que articula preprocesamiento, modelado, validación cuantitativa y análisis discursivo complementario.

Finalmente, el preprocesamiento se ajusta al tipo de modelo. En LDA se aplica normalización léxica y lematización para reducir variación morfológica, mientras que en BERTopic se utiliza una limpieza básica, dado que la representación semántica se deriva principalmente de embeddings contextuales. Esta estrategia permite mantener comparabilidad con la línea base y aprovechar las ventajas del enfoque contextual.

4.8. Sentimiento, polaridad y tono

El análisis de sentimiento es una de las aplicaciones más consolidadas del Procesamiento de Lenguaje Natural en estudios de comunicación política y mediática. Este enfoque permite estimar la orientación evaluativa de los textos, clasificándolos según su polaridad (positiva, negativa o neutral) y cuantificando la intensidad de dicha valoración. En el periodismo digital, estas dimensiones resultan fundamentales para analizar cómo los medios construyen representaciones emocionales sobre actores e instituciones (Liu, 2020).

Los modelos actuales de análisis de sentimiento utilizan representaciones contextuales del lenguaje derivadas de arquitecturas basadas en transformer (Vaswani et al., 2017). Modelos como BERT y sus variantes permiten capturar dependencias semánticas complejas y mejorar la clasificación de polaridad frente a enfoques léxicos tradicionales (Devlin et al., 2019). En el análisis mediático, estos modelos han demostrado mayor capacidad para identificar matices evaluativos en textos periodísticos y políticos (Zhang et al., 2023).

En el contexto de la corrupción en salud, la estimación de polaridad permite detectar patrones de negatividad, diferencias emocionales entre medios y variaciones temporales del tono discursivo. Diversos estudios indican que la cobertura de corrupción tiende a enfatizar narrativas de indignación moral y desconfianza institucional, lo que se refleja en una predominancia de polaridad negativa (Arroyave & Barrios, 2023; Fan et al., 2022).

En esta investigación, la polaridad se define como la clasificación del texto en positivo, neutral o negativo, mientras que el tono se aproxima mediante la intensidad probabilística asociada a dicha clasificación. En este sentido, el término tono no se emplea como inferencia sobre la intención del emisor, sino como una aproximación a la carga emocional o evaluativa predominante del discurso mediático. El análisis adopta un enfoque comparativo que combina un método léxico como línea base con modelos contextuales en español como estrategia

principal, permitiendo evaluar la consistencia entre enfoques y fortalecer la robustez del análisis emocional.

Los resultados se presentan como distribuciones de polaridad por medio y periodo (2022–2023), promedios de negatividad por tópico y evolución temporal del tono, con el fin de examinar la coherencia entre las estructuras temáticas y la carga evaluativa del discurso mediático.

4.9. Métricas y validación

La evaluación de los modelos de Procesamiento de Lenguaje Natural empleados en este estudio se fundamenta en métricas orientadas a estimar coherencia semántica, diversidad temática y estabilidad estructural de las particiones generadas. Dado el carácter no supervisado del modelado temático, la validación se centró en medidas internas y en la concordancia entre modelos, en lugar de métricas propias de clasificación supervisada.

En el modelado temático, la selección del modelo principal se realizó mediante la métrica de coherencia C_v, reconocida por su correlación con la interpretabilidad humana de los tópicos (Röder et al., 2015). Esta medida se complementó con C_npmi, basada en información mutua normalizada entre términos (Bouma, 2009; Lau et al., 2014), y con Topic Diversity (TD), que evalúa la diferenciación léxica entre temas (Dieng et al., 2020). La combinación de estas métricas permitió identificar soluciones equilibradas entre cohesión semántica y separación conceptual.

Para evaluar la estabilidad estructural, se compararon las particiones obtenidas por LDA y BERTopic mediante NMI (Normalized Mutual Information) y ARI (Adjusted Rand Index), métricas ampliamente utilizadas para contrastar agrupamientos no supervisados (Hubert & Arabie, 1985; Strehl & Ghosh, 2003; Vinh et al., 2010). Estas medidas permiten estimar el grado de convergencia entre asignaciones por documento sin requerir etiquetas externas. Asimismo, el modelo HDP se utilizó como contraste no paramétrico para explorar el rango

plausible de tópicos. De forma complementaria, se evaluó la similitud léxica entre los tópicos obtenidos por los distintos modelos mediante el índice de Jaccard aplicado a los términos principales (top-words), lo que permitió estimar el grado de solapamiento léxico-semántico entre las representaciones temáticas.

En el análisis de sentimiento, se empleó un modelo contextual basado en Transformers como estimador principal de polaridad emocional. El modelo asigna a cada documento probabilidades asociadas a las categorías de polaridad (positiva, neutra y negativa), utilizándose la probabilidad de negatividad como indicador central del tono discursivo.

La evaluación se sustentó en la consistencia de las distribuciones del tono emocional entre medios, ejes temáticos y periodos, así como en la convergencia de resultados con recursos léxicos especializados utilizados como referencia metodológica. Este enfoque permitió identificar patrones diferenciales del discurso sin asumir supuestos de normalidad ni requerir pruebas inferenciales paramétricas.

En conjunto, la estrategia de validación integra coherencia temática, diversidad léxica, estabilidad intermodelo y análisis estadístico del componente emocional, configurando un marco robusto y reproducible para el estudio del discurso mediático.

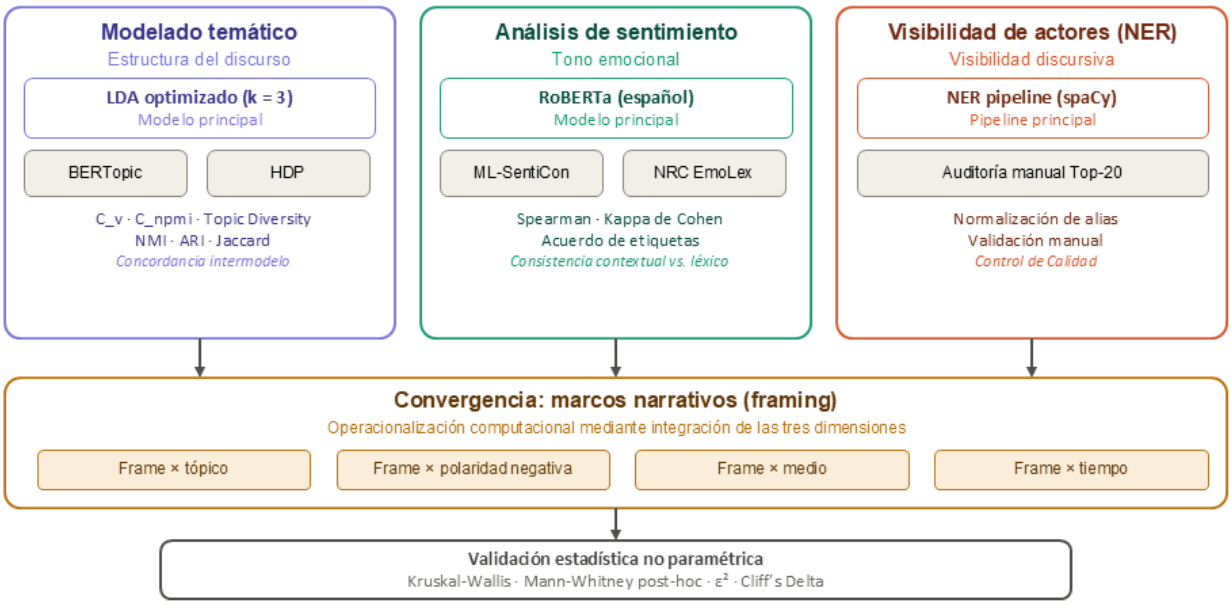
Tabla 2. Matriz de triangulación y comparación de modelos por componente

Componente	Tópicos (estructura temática)	Sentimiento (polaridad/tono)	Actores (visibilidad)	Frames / marcos narrativos (síntesis discursiva)
Modelo principal	**LDA optimizado**	**RoBERTa entrenado para español (modelo contextual)**	**NER pipeline**	**Esquema de frames + scoring (proxy)**
Modelos de contraste	BERTopic (contextual) + HDP	Baselines léxicos ML-SentiCon + NRC EmoLex (contraste metodológico)	Auditoría manual Top-20 (control de calidad)	Tópicos + sentimiento (como validación convergente)
Pares comparativos	LDA↔BERTopic HDP → LDA (plausibilidad de k)	RoBERTa ↔ ML-SentiCon RoBERTa ↔ NRC EmoLex	NER↔Auditoría	Frame x Tópico Frame × polaridad negativa Frame x Medio/tiempo
Qué se contrasta	Robustez/estabilidad temática bajo supuestos distintos	Consistencia direccional (tendencias) por medio/tiempo/tópico	Calidad de extracción (ruido/alias) y estabilidad del top	Si el frame aporta capa interpretativa distinta a "tópico renombrado"
Métricas/criterios mínimos	C_v + C_npmi + Topic Diversity (TD) NMI/ARI para particiones Jaccard (Top-words, similitud léxica) % de documentos	Correlación documental (Spearman); consistencia de ranking por medio y tópico; kappa de Cohen	Overlap Top-20; revisión de falsos positivos/alias en muestra pequeña	Asociación (topic×frame) + verificación interpretativa mínima (2–3 ejemplos por frame, verificación interpretativa); estabilidad temporal básica

	asignados a ruido (-1) en BERTopic # tópicos "activos" en HDP			
Output para reportar	(i) k final (x) + justificación (ii) consistencia LDA↔BERTopic (iii) rango HDP (tópicos activos)	Distribución de polaridad; promedio de negatividad por medio/tópico/período; evidencia de robustez mediante comparación entre modelo contextual y baselines léxicos	Top-20 actores depurado + nota de control de calidad	3–4 frames máximos, interpretados como aproximación operacional a framing

Nota. La tabla sintetiza la estrategia de triangulación intermodelo y los criterios de contraste utilizados para evaluar estabilidad temática, consistencia emocional y calidad de extracción de actores.

Figura 2. Esquema de triangulación metodológica del sistema computacional de análisis discursivo



Nota. La figura sintetiza la estrategia de triangulación metodológica del estudio, integrando los componentes de modelado temático, análisis de sentimiento y visibilidad de actores, así como sus mecanismos de contraste, convergencia analítica y validación estadística no paramétrica.

4.10. Operacionalización y trazabilidad

La operacionalización traduce los conceptos teóricos y analíticos del estudio corrupción, discurso mediático, framing y tono emocional en variables observables y medibles dentro del corpus periodístico. Este proceso garantiza coherencia entre el marco conceptual y la ejecución

técnica del análisis, permitiendo que los resultados sean reproducibles y verificables (Adcock & Collier, 2016).

Esta traducción se implementa mediante un esquema de trazabilidad que conecta cada nivel conceptual con los procedimientos computacionales aplicados, articulando dimensiones discursivas, lingüístico-computacionales, técnicas y de validación (Tabla 3).

Tabla 3. Dimensiones analíticas y su función en el modelo discursivo

Dimensión	Descripción / Componentes	Propósito analítico
Discursiva	Marcos narrativos, atribución de responsabilidad, relaciones entre actores, configuración interpretativa del discurso.	Interpretar patrones discursivos y marcos de significado a partir de la integración de resultados temáticos, emocionales y actorales, en coherencia con la teoría del framing mediático.
Lingüística-computacional	Tópicos temáticos, polaridad y tono emocional, distribución léxica y patrones textuales derivados de modelos de PLN.	Caracterizar cuantitativamente el contenido noticioso mediante variables textuales obtenidas a partir de modelado temático y análisis de sentimiento.
Técnica	Recolección automatizada, limpieza, normalización, lematización, tokenización y construcción de representaciones textuales adecuadas para los modelos aplicados.	Garantizar trazabilidad metodológica y consistencia en la preparación del corpus para el análisis computacional.
Validación	Métricas de coherencia temática (C_v, C_npmi), diversidad temática (Topic Diversity), estabilidad intermodelo (NMI, ARI) y similitud léxica (Jaccard).	Evaluar la consistencia interna, robustez estructural y reproducibilidad del modelado temático en ausencia de etiquetas externas.

Nota. La tabla sintetiza las dimensiones conceptuales, computacionales y técnicas que estructuran el análisis discursivo mediante PLN, integrando componentes teóricos, variables textuales y criterios de validación empleados en el estudio.

Con base en estas dimensiones, la evaluación de resultados se reporta mediante comparaciones entre modelos de tópicos (LDA y BERTopic) utilizando métricas de coherencia y diversidad temática, así como medidas de estabilidad cuando se contrastan asignaciones entre modelos. En el componente de sentimiento, dado que no se dispone de etiquetado manual, la validación se aborda mediante triangulación entre un método léxico y un modelo contextual en español, evaluando concordancia mediante correlación de rangos (Spearman) y estabilidad direccional por medio y periodo.

El principio de trazabilidad se fundamenta en los lineamientos de reproducibilidad científica y open science (Peng, 2011; Stodden et al., 2018), según los cuales los análisis basados en datos deben poder replicarse bajo las mismas condiciones técnicas. Para garantizarlo, el estudio documenta sistemáticamente las fuentes de datos, las versiones del código del pipeline, los parámetros de configuración y los resultados intermedios necesarios para verificación externa.

Asimismo, se define un conjunto de variables operativas que describen tanto las transformaciones del texto como las salidas generadas por el pipeline de PLN, permitiendo rastrear cada resultado hasta su origen empírico y computacional (Tabla 4).

Tabla 4. Variables analíticas y métricas derivadas del pipeline de PLN aplicado al corpus

Variable final	Qué es	Cómo se calcula	Etapa	Para qué sirve	Cómo se interpreta
ID_Noticia	ID único por noticia	hash o consecutivo al capturar	Recolección/BD	Trazabilidad	Rastrear resultados al texto original
Medio	Fuente del artículo	del sitio	Recolección	Comparar medios	Patrones editoriales
Fecha	Publicación	HTML/metadatos	Recolección	Temporal 2022–2023	Cambios en el tiempo
URL	Enlace origen	almacenamiento del enlace	Recolección	Verificación	Reproducibilidad
Cobertura (nacional/regional)	Tipo de medio	clasificación del medio	Metadatos	Comparar cobertura	Contraste nacional vs regional
Texto_original	Texto base	titular + cuerpo	Corpus	Insumo principal	Discurso “real”
Texto_preprocesado	Texto normalizado	limpieza + normalización (LDA: incluye lematización; BERTopic: limpieza básica)	Preprocesamiento	Mejorar tópicos	Menos ruido
Longitud_documento	Tamaño del texto	word_count / char_count / length_category	Preprocesamiento	Control de sesgos	Corto vs largo
Tópico_LDA	Tema por LDA (línea base)	tópico dominante	Modelado	Línea base	Comparación/interpretación
Peso_tópico_dominante	Fuerza del tópico	peso del tópico dominante	Modelado	Intensidad temática	Más alto = más representativo
Top_palabras_tópico	Top palabras por tópico	Top-N palabras con mayor probabilidad	Modelado	Interpretar tópicos	“De qué trata”
Tópicos_por_medio	Distribución por medio	% por tópico y medio	Agregación	Agenda/encuadre	Diferencias entre medios
Tópicos_por_periodo	Distribución temporal	% por mes/trimestre/año	Agregación	Evolución	Cambios discursivos
Polaridad_lexica	Polaridad baseline	modelo léxico	Sentimiento	Triangulación	Comparación rápida

Polaridad_contextual	Polaridad principal	Modelo transformer en español	Sentimiento	Resultado principal	Capta matices
Intensidad_polaridad	Intensidad	puntuación continua u ordinal	Sentimiento	Tono como intensidad	Más negativo = más carga
Entidades_NER	Entidades nombradas detectadas	modelo NER + normalización de alias	Actores	Identificar protagonistas	Actores más visibles
Actor_principal	Entidad más frecuente por noticia	conteo por documento	Actores	Identificar actor predominante	Actor dominante
Top_actores	Ranking de actores del corpus	conteo agregado	Agregación	Identificar figuras clave	Jerarquía actoral
Frame_dominante	Marco narrativo principal	scoring tópico-sentimiento-actor	Integración discursiva	Síntesis interpretativa	Tipo de encuadre dominante
Frames_por_medio	Distribución de marcos por medio	% por frame y medio	Agregación	Comparar narrativas editoriales	Diferencias discursivas
C_v	Coherencia tópicos	métrica C_v	Validación	Calidad tópicos	Más alto = mejor
C_npmi	Coherencia normalizada	métrica C_npmi	Validación	Robustez	Más alto = mejor
Topic_Diversity	Diversidad tópicos	proporción palabras únicas	Validación	Evitar repetición	Más alto = menos redundancia
NMI	Concordancia LDA–BERTopic	métrica NMI	Validación	Estabilidad estructural	Más alto = mayor acuerdo
ARI	Concordancia ajustada	métrica ARI	Validación	Robustez	Más alto = mayor acuerdo
Jaccard_top_words	Similitud léxica entre tópicos	intersección/unión top-words	Validación	Consistencia semántica	Más alto = mayor similitud
Docs_ruido_BERTopic	% documentos sin asignación	conteo label −1	Validación	Control de clustering	Más alto = más ruido, Indica proporción de documentos no agrupados
Topicos_activos_HDP	Número de tópicos estimados	salida HDP	Validación	Rango plausible de k	Comparación con LDA

Nota. Las variables se organizan según las etapas del pipeline: recolección, preprocesamiento, modelado temático, análisis de sentimiento, extracción de entidades, integración discursiva y validación. Las métricas C_v, C_npmi y Topic Diversity evalúan la calidad interna del modelado temático, mientras que NMI, ARI y Jaccard estiman la concordancia entre modelos no supervisados.

La trazabilidad metodológica asegura que los resultados puedan vincularse con los datos originales y con las decisiones analíticas adoptadas, fortaleciendo la transparencia, la reproducibilidad y la integridad científica del estudio. De este modo, el marco conceptual y la metodología se integran en un sistema coherente donde cada componente del análisis es documentado y verificable.

4.11. Vacíos, riesgos y sesgos

Todo análisis automatizado del discurso implica limitaciones inherentes y riesgos epistemológicos que deben reconocerse explícitamente para mantener la transparencia científica. En el estudio de la corrupción en salud mediante PLN, los principales desafíos derivan tanto del comportamiento de los modelos lingüísticos como de las características del corpus periodístico.

Uno de los riesgos más relevantes es el sesgo de representación mediática. Los medios digitales tienden a priorizar narrativas de escándalo, confrontación política y denuncia moral, lo que puede sobrerrepresentar ciertos actores o regiones y subestimar otros aspectos estructurales del fenómeno (Arroyave & Barrios, 2023; Tandoc et al., 2022). Este sesgo afecta la diversidad temática del corpus y puede influir en la interpretación de los resultados, reforzando marcos de atribución moral o institucional.

Desde el punto de vista técnico, los modelos de PLN incorporan sesgos algorítmicos derivados de los datos con los que fueron entrenados. Investigaciones recientes demuestran que incluso los modelos preentrenados como BERT o la familia GPT reproducen patrones de desigualdad o estereotipos lingüísticos presentes en sus datos fuente (Bender et al., 2021; Blodgett et al., 2020). Esto obliga a aplicar estrategias de control analítico, como la triangulación intermodelo, la comparación de resultados entre enfoques probabilísticos y contextuales y la interpretación crítica de los tópicos y polaridades detectadas.

Otro límite importante es el tamaño y naturaleza de los textos analizados. Los titulares o notas breves presentan baja densidad semántica, lo que puede reducir la coherencia de los modelos temáticos (Röder et al., 2015). Para mitigar esta limitación, se incorporaron controles técnicos de calidad en el preprocesamiento, incluyendo eliminación de duplicados, limpieza de boilerplate, modelado de bigramas y trigramas para preservar unidades semánticas compuestas, y exclusión de textos residuales tras limpieza. Estos controles no buscan excluir

noticias por pertenecer a un “tópico”, sino asegurar consistencia técnica y comparabilidad del corpus.

A nivel metodológico, el estudio reconoce que la automatización no sustituye la interpretación humana. Los resultados derivados del PLN deben entenderse como insumos empíricos complementarios que enriquecen, pero no reemplazan, la lectura crítica del discurso (Jacobs & Tschötschel, 2019). Este principio mantiene la conexión entre el rigor computacional y la reflexión teórica, evitando que los modelos se conviertan en cajas negras interpretativas.

Aunque no se realizó una validación léxica específica sobre colombianismos o regionalismos del discurso de corrupción, la aptitud de los modelos se sustenta en su entrenamiento para español, su capacidad contextual, el uso de n-gramas para preservar expresiones compuestas del dominio y la triangulación metodológica aplicada en el estudio.

Finalmente, se asume como límite estructural la dinámica evolutiva del discurso digital. La constante actualización de los medios, cambios en formatos narrativos y políticas editoriales pueden alterar la distribución temática a lo largo del tiempo, afectando la replicabilidad longitudinal del análisis. Reconocer estos vacíos y sesgos es esencial para contextualizar los hallazgos y fortalecer la validez interna y externa del estudio.

4.12. Síntesis integradora

El presente marco teórico articula un recorrido conceptual que vincula el fenómeno social de la corrupción en salud, su construcción mediática digital y las técnicas de Procesamiento de Lenguaje Natural (PLN) empleadas para su análisis. Esta integración establece la base epistemológica y metodológica del estudio, garantizando coherencia entre los objetivos, el enfoque teórico y las herramientas computacionales utilizadas.

En primer lugar, la corrupción en salud fue abordada como un problema estructural de gobernanza y legitimidad institucional, cuya representación mediática configura percepciones sociales sobre transparencia y eficiencia estatal. Los medios digitales actúan como agentes

discursivos que contribuyen a la construcción de marcos interpretativos del fenómeno, en función del sesgo editorial y del tipo de cobertura (Arroyave & Barrios, 2023).

Desde la teoría del discurso y el framing, se comprendió que las estructuras lingüísticas y narrativas reproducen relaciones de poder y orientan la interpretación del público (Entman, 1993; Fairclough, 2013). Este fundamento conceptual dio sustento a la aplicación del PLN como herramienta para operacionalizar el discurso mediático en unidades cuantificables, sin perder su dimensión semántica y crítica.

El conjunto de técnicas consideradas modelado temático, análisis de sentimiento y validación métrica, configura un enfoque computacional interpretativo y replicable, en el que el análisis automatizado complementa la reflexión teórica. En particular, LDA como línea base y BERTopic junto con HDP como enfoques comparativos permiten identificar núcleos temáticos del discurso; los modelos basados en Transformer capturan dimensiones emocionales del lenguaje; y la triangulación intermodelo fortalece la robustez estructural de los hallazgos.

La validez analítica se sustenta en métricas cuantificables de coherencia temática (C_v, C_npmi y Topic Diversity) y en medidas de estabilidad estructural entre modelos no supervisados (NMI, ARI y similitud léxica Jaccard). En ausencia de etiquetas manuales externas, la consistencia de los resultados se evaluó mediante triangulación metodológica entre modelos probabilísticos, contextuales y no paramétricos.

En el componente de sentimiento, dado que no se empleó etiquetado manual, la robustez se estimó mediante la convergencia entre un método léxico de línea base y un modelo Transformer en español, reportando el grado de acuerdo entre ambos enfoques.

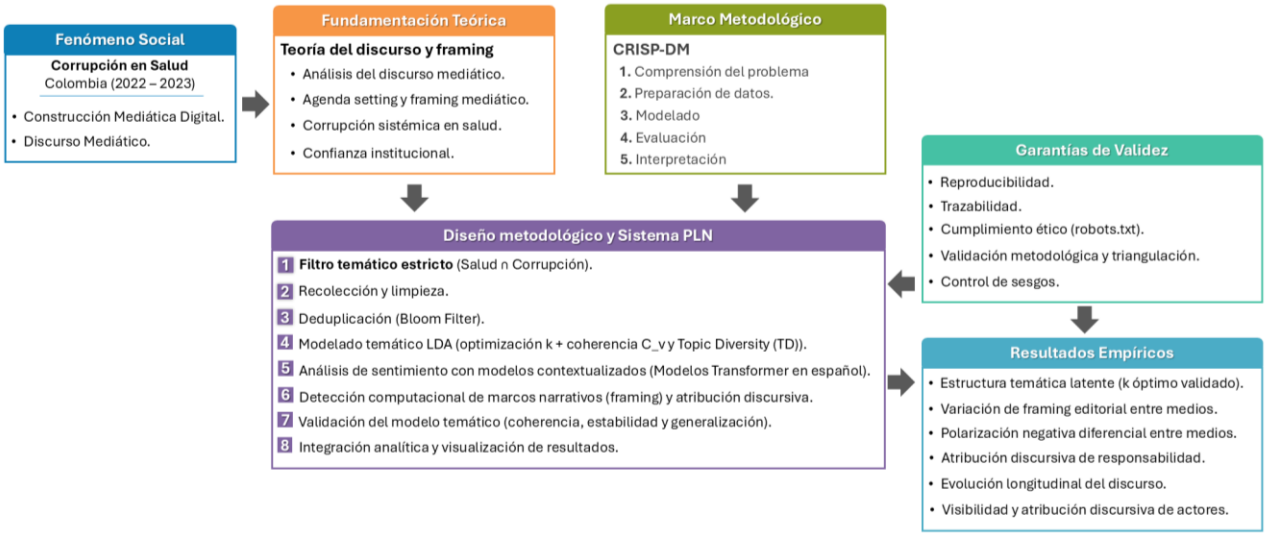
La trazabilidad metodológica, alineada con principios de reproducibilidad científica, permite vincular cada resultado con su procedimiento técnico y fundamento conceptual (Peng, 2011; Stodden et al., 2018). Asimismo, el reconocimiento explícito de sesgos y limitaciones refuerza la integridad analítica del estudio y la necesidad de una interpretación crítica complementaria.

En conjunto, el marco teórico consolida una articulación integrada entre fenómeno, teoría, método y métrica, en la que la ciencia de datos se pone al servicio del análisis crítico del discurso mediante una arquitectura de triangulación algorítmica que contrasta modelos probabilísticos, contextuales y no paramétricos. Esta arquitectura conceptual habilita la fase empírica del proyecto, en la que los modelos de PLN permitieron examinar cómo los medios digitales colombianos construyen y difunden las narrativas sobre corrupción en salud durante el periodo 2022–2023, evaluando su estructura temática, carga emocional y estabilidad intermodelo.

En este sentido, las técnicas computacionales empleadas en el estudio no constituyen un Análisis Crítico del Discurso en sentido hermenéutico estricto, sino una estrategia de operacionalización empírica de regularidades discursivas que requiere interpretación teórica posterior. Por ello, el ACD se adopta como marco de referencia para la lectura sustantiva de los hallazgos, mientras que el PLN aporta insumos cuantificables y replicables que complementan, pero no reemplazan, la interpretación cualitativa clásica.

En síntesis, el marco teórico no solo sustenta conceptualmente el estudio, sino que orienta la selección y aplicación de cada técnica de análisis: el framing para caracterizar los encuadres mediáticos, el PLN para modelar patrones lingüísticos y emocionales, y las métricas de coherencia para respaldar la validez cuantitativa de los resultados, asegurando así una relación directa entre teoría, método y objetivos.

Figura 3. Mapa conceptual integrador del diseño teórico-metodológico y resultados



Nota. El esquema sintetiza la articulación entre el fenómeno social, la fundamentación teórica, el marco metodológico (CRISP-DM), el diseño operacional del sistema PLN y los resultados empíricos, integrados bajo criterios de validez, trazabilidad y evaluación estadística.

5. Hipótesis

El presente proyecto se orienta por una hipótesis nula (H_0) y una hipótesis alternativa (H_1), formuladas con base en el objetivo general de identificar patrones lingüísticos y temáticos en los discursos sobre corrupción en el sector salud en Colombia (2022–2023), publicados en medios de comunicación digitales, mediante técnicas de Procesamiento de Lenguaje Natural (PLN).

Las hipótesis se formulan en términos de detectabilidad empírica mediante técnicas de PLN no supervisado y análisis estadístico de las variables derivadas del corpus.

Estas hipótesis se expresan de la siguiente manera:

5.1. Hipótesis Nula (H_0)

“No existen estructuras temáticas coherentes ni diferencias estadísticamente significativas en la distribución del tono emocional del discurso sobre corrupción en el sector salud en Colombia (2022–2023) entre los medios digitales analizados, evaluadas mediante métricas de coherencia temática del modelado LDA y los contrastes no paramétricos aplicados a las medidas de polaridad”.

5.2. Hipótesis Alternativa (H_1)

“Existen estructuras temáticas coherentes y diferencias estadísticamente significativas en la distribución del tono emocional del discurso sobre corrupción en el sector salud en Colombia (2022–2023) entre los medios digitales analizados, identificables mediante modelado temático LDA validado por métricas de coherencia semántica y análisis de sentimiento evaluado mediante contrastes no paramétricos”.

La contrastación formal de H_0 y H_1 se sustenta en dos niveles complementarios: (i) la validación del modelado temático mediante métricas de coherencia semántica y diversidad temática, y (ii) la evaluación de diferencias en la distribución del tono emocional entre grupos

discursivos mediante pruebas estadísticas no paramétricas. Los resultados formales de estas pruebas se presentan en la subsección correspondiente del capítulo de resultados, donde se reportan el nombre de la prueba, el estadístico, el valor p y el tamaño del efecto para cada contraste relevante.

6. Variables

Las variables derivadas corresponden a indicadores computacionales del discurso, no a mediciones directas del fenómeno social.

En la presente investigación, se identifican las siguientes variables clave, con su respectiva definición conceptual, operacional y clasificación correspondiente:

6.1. Discurso sobre corrupción en salud

Tipo: Variable dependiente principal.

Definición conceptual

Construcción discursiva mediante la cual los medios digitales representan hechos, actores y dinámicas asociadas a corrupción en el sistema de salud colombiano durante 2022–2023.

Definición operacional

Se operacionaliza a partir de un corpus final de 518 documentos, seleccionados bajo intersección estricta SALUD ∩ CORRUPCIÓN, depurados mediante:

- Limpieza estructural.
- Eliminación de boilerplate.
- Deduplicación con Bloom Filter.
- Validación temática estricta.
- Exclusión de medios con robots.txt restringido.

El discurso se modela computacionalmente mediante:

- Modelado temático LDA validado (C_v, C_npmi y Topic Diversity).
- Análisis de sentimiento probabilístico.
- Aproximación computacional a marcos narrativos (framing) a partir de la convergencia entre tópicos, tono emocional y visibilidad de actores.
- Identificación de atribución de responsabilidad mediante análisis de actores y contextos temáticos.
- Visibilidad de actores (NER).

Clasificación

Constructo analítico central operacionalizado mediante variables latentes derivadas (tópicos, tono, frames y actores).

6.2. Tipo de medio digital de noticias

Tipo: Variable independiente estructural.

Definición conceptual

Categoría editorial según alcance territorial del medio.

Definición operacional

Clasificación dicotómica:

- Medio nacional (n = 4).
- Medio regional (n = 2).

Clasificación

Variable cualitativa nominal manifiesta.

6.3. Tono del discurso

Tipo: Variable analítica derivada del discurso.

Definición conceptual

Carga emocional predominante asociada al tratamiento mediático del fenómeno de corrupción en el sector salud.

Definición operacional

Estimación del tono emocional mediante un modelo supervisado preentrenado basado en Transformers en español, capaz de capturar relaciones semánticas contextuales. El modelo asigna a cada documento probabilidades asociadas a las categorías de polaridad (positiva, neutra y negativa), para el análisis se utilizaron:

- Probabilidad de polaridad negativa como indicador principal del tono emocional.

- Clasificación discreta derivada (positivo / neutro / negativo) según la categoría con mayor probabilidad.

El análisis se complementa con recursos léxicos especializados como referencia metodológica adicional, con el fin de evaluar la consistencia de los resultados entre enfoques contextuales y basados en diccionarios.

Clasificación

Variable cuantitativa continua (probabilidad de negatividad) con categorización derivada ordinal.

Consideración

El estudio adopta un enfoque descriptivo y comparativo del tono emocional entre medios y ejes temáticos, priorizando la interpretación de distribuciones y patrones relativos en lugar de pruebas inferenciales paramétricas, dado el carácter heterogéneo y contextual del discurso mediático. La triangulación entre modelos supervisados y recursos léxicos permite reforzar la robustez interpretativa de los resultados.

6.4. Tópicos discursivos sobre corrupción en salud

Tipo: Variable analítica latente derivada.

Definición conceptual

Estructuras semánticas latentes que organizan el discurso mediático.

Definición operacional

Identificación mediante modelo LDA optimizado bajo:

- Coherencia semántica C_v.
- Coherencia normalizada C_npmi.
- Topic Diversity.

Cada documento recibe:

- Distribución probabilística por tópico.

- Tópico dominante.

Clasificación

Variable cualitativa latente con representación probabilística continua.

Consideración

La consistencia estructural de los tópicos se valida mediante métricas de coherencia semántica (C_v y C_npmi) y diversidad temática (Topic Diversity), las cuales permiten evaluar la cohesión interna, la robustez léxica y la diferenciación conceptual del modelo LDA seleccionado.

Asimismo, la plausibilidad de la estructura temática se contrastó con modelos alternativos no supervisados, sin alterar la selección del modelo principal.

6.5. Frecuencia de términos clave

Tipo: Variable descriptiva complementaria.

Definición conceptual

Intensidad léxica asociada a vocabulario temático, se utiliza con fines descriptivos e interpretativos, no inferenciales.

Definición operacional

Cálculo mediante:

- TF normalizado. (Discreta)
- TF-IDF descriptivo. (Continua)

Se utiliza para:

- Interpretación semántica de tópicos.
- Validación léxica de frames.

Clasificación

Variable cuantitativa de razón (continua y discreta según el indicador aplicado).

6.6. Temporalidad de publicación

Tipo: Variable independiente contextual.

Definición conceptual

Momento cronológico de publicación.

Definición operacional

Fecha extraída automáticamente y agrupada en:

- Año.
- Mes.

Para análisis de:

- Evolución temática.
- Evolución del sentimiento.
- Intensidad de frames.

Clasificación

Variable temporal discreta tratada como ordinal para análisis comparativo y como serie temporal agregada para análisis longitudinal.

6.7. Entidad mencionada en la noticia

Tipo: Variable analítica derivada.

Definición conceptual

Actor institucional o individual visibilizado en el discurso, control de alias y normalización manual para entidades de alta frecuencia.

Definición operacional

Identificación automática mediante NER en español.

Posterior normalización y agregación en:

- Entidades públicas.
- Entidades privadas.
- Personas.

Se calculan:

- Frecuencia global.
- Distribución por medio.
- Co-ocurrencia en red.

Clasificación

Variable cualitativa nominal con conteo cuantitativo derivado.

7. Metodología

Con base en el marco teórico desarrollado, las hipótesis formuladas y la operacionalización de variables presentada en el capítulo anterior, el presente acápite describe el diseño metodológico adoptado para evaluar empíricamente la presencia de estructuras temáticas consistentes y diferencias estadísticamente significativas en las medidas de polaridad emocional del discurso sobre corrupción en salud en Colombia (2022–2023).

La metodología integra procedimientos de recolección automatizada, preprocesamiento y modelado computacional de un corpus noticioso digital, articulando técnicas de Procesamiento de Lenguaje Natural (PLN) con inferencia estadística no paramétrica aplicada a comparaciones entre grupos discursivos. Se detallan el enfoque, diseño, alcance y fases del estudio, así como la unidad de análisis, la muestra seleccionada, los instrumentos tecnológicos empleados y las métricas de validación aplicadas, garantizando coherencia entre fundamentos teóricos, hipótesis, variables y procedimientos analíticos.

La validación del componente de modelado temático se realizó mediante un esquema de triangulación metodológica propio de enfoques no supervisados, estructurado en tres niveles complementarios:

- Validación interna del modelo, basada en métricas de coherencia semántica (C_v y C_npmi) y diversidad temática (Topic Diversity), utilizadas para evaluar la calidad interpretativa, la cohesión semántica y la diferenciación entre tópicos.
- Contraste intermodelo, mediante la aplicación de enfoques alternativos de modelado temático (HDP y BERTopic), con el propósito de examinar la estabilidad estructural del espacio temático identificado y la plausibilidad del número de tópicos seleccionado.

- Validación estadística y analítica, que incluyó pruebas no paramétricas de comparación entre grupos discursivos (por medio, periodo y tópico), análisis cruzado de variables temáticas, emocionales y actorales.

Este esquema permite evaluar simultáneamente la calidad interna del modelo, su robustez frente a aproximaciones alternativas y su consistencia empírica dentro del análisis del discurso mediático. La validación se orienta a consistencia, interpretabilidad y estabilidad estructural, no a exactitud predictiva.

7.1. Enfoque de investigación

La presente investigación adopta un enfoque cuantitativo de carácter observacional, no experimental y retrospectivo, fundamentado en el análisis computacional de datos textuales mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Este enfoque permite transformar estructuras lingüísticas en representaciones numéricas susceptibles de medición, comparación e inferencia estadística, práctica ampliamente consolidada en la minería de texto aplicada a las ciencias sociales (Camacho-Collados & Pilehvar, 2018; Grimmer & Stewart, 2013).

Desde el punto de vista analítico, la investigación se estructura como un estudio comparativo entre grupos independientes de medios digitales nacionales y regionales, con el fin de evaluar diferencias en la distribución del tono emocional del discurso y en la configuración temática latente. La contrastación de hipótesis se realiza mediante pruebas estadísticas no paramétricas, adecuadas para variables derivadas de datos textuales que no cumplen supuestos de normalidad.

El análisis temático se implementa mediante Latent Dirichlet Allocation (LDA), modelo probabilístico generativo propuesto por Blei et al. (2003), cuya consistencia estructural se valida a través de métricas de coherencia semántica (C_v y C_npmi) y diversidad temática (Topic

Diversity) (Röder et al., 2015). Este procedimiento permite identificar estructuras latentes recurrentes en el corpus sin intervención supervisada.

De manera complementaria, el tono emocional se estima mediante un modelo supervisado de análisis de sentimiento basado en representaciones contextuales del lenguaje, previamente entrenado para español y aplicado al corpus sin ajuste específico (Devlin et al., 2019; Yin et al., 2021). La salida del modelo se expresa como un score continuo de polaridad emocional en escala [-1,1], lo que permite su tratamiento como variable cuantitativa continua para la comparación entre grupos mediante técnicas no paramétricas.

En consecuencia, el estudio articula modelado no supervisado, clasificación supervisada e inferencia estadística, integrando análisis semántico automatizado con validación cuantitativa, en coherencia con el marco conceptual del discurso y la aproximación computacional al framing mediático desarrollada en capítulos anteriores.

7.2. Diseño de investigación

La investigación se desarrolla bajo un diseño no experimental, transversal y comparativo de carácter retrospectivo. Se considera no experimental debido a que las variables analizadas no son manipuladas, sino observadas en su contexto natural de producción discursiva, a partir de contenidos publicados por medios digitales durante el periodo 2022–2023 (Creswell, 2014; Hernández-Sampieri et al., 2018).

El carácter transversal se fundamenta en que el análisis se realiza sobre un corpus histórico delimitado temporalmente, sin seguimiento prospectivo ni mediciones repetidas sobre las unidades de análisis, aunque incorporando análisis de variación dentro del intervalo temporal considerado. Este tipo de diseño resulta adecuado cuando se estudian fenómenos sociales ya ocurridos a partir de datos documentales o registros históricos (Creswell, 2014).

El diseño es comparativo entre grupos independientes, dado que se contrastan diferencias en la estructura temática y en la distribución del tono emocional entre medios

digitales nacionales y regionales. Esta aproximación permite identificar variaciones sistemáticas entre categorías preexistentes sin intervención experimental (Hernández-Sampieri et al., 2018).

La unidad de análisis corresponde al documento periodístico individual, tratado como observación analítica dentro del corpus, mientras que el nivel de agregación secundaria se establece a nivel de medio digital, el cual actúa como variable agrupadora para el análisis comparativo. Esto permite evaluar patrones discursivos tanto a nivel de documentos individuales como de tendencias por medio.

En consecuencia, el diseño adoptado permite describir, modelar y contrastar empíricamente patrones discursivos sin manipulación de variables, garantizando coherencia metodológica con las hipótesis planteadas y las variables operacionalizadas.

7.3. Alcance de la investigación

El estudio presenta un alcance descriptivo, comparativo e inferencial, con componentes exploratorios derivados del uso de técnicas automatizadas de análisis textual.

Es descriptivo en la medida en que caracteriza la estructura temática y la distribución del tono emocional del discurso sobre corrupción en salud emitido por medios digitales durante el periodo 2022–2023, permitiendo identificar regularidades semánticas y narrativas en el corpus analizado.

Es comparativo porque contrasta dichas estructuras entre grupos independientes de medios digitales nacionales y regionales con el propósito de examinar variaciones sistemáticas en la configuración discursiva.

Asimismo, el estudio incorpora un componente inferencial al evaluar la existencia de diferencias estadísticamente significativas en la distribución del tono emocional mediante pruebas no paramétricas, lo que permite contrastar empíricamente las hipótesis planteadas sin establecer relaciones causales, dado el carácter observacional del diseño.

El carácter exploratorio se circunscribe específicamente al modelado temático no supervisado (LDA), técnica que permite identificar estructuras semánticas latentes sin categorías predefinidas, en coherencia con enfoques emergentes de análisis automatizado de contenido en comunicación digital (Boumans & Trilling, 2016).

En conjunto, el alcance adoptado permite describir, modelar y contrastar patrones discursivos mediante procedimientos reproducibles, triangulación analítica y fundamentación estadística, en coherencia con los objetivos y el diseño metodológico del estudio.

7.4. Tipo de investigación

La presente investigación es de carácter aplicado, ya que desarrolla e implementa una metodología analítica reproducible orientada a la caracterización empírica del discurso mediático digital sobre corrupción en el sector salud en Colombia. Su propósito no se limita a la generación de conocimiento teórico, sino que busca aportar un procedimiento sistemático transferible para el análisis de fenómenos discursivos mediante técnicas computacionales.

Se inscribe en la investigación cuantitativa aplicada en ciencias sociales computacionales, al emplear técnicas de Procesamiento de Lenguaje Natural (PLN), modelado temático probabilístico y análisis supervisado de sentimiento para el tratamiento automatizado de grandes volúmenes de texto. Este tipo de aproximación ha sido ampliamente respaldado en la literatura reciente como un mecanismo válido para el análisis sistemático de fenómenos sociales mediados por comunicación digital (Boumans & Trilling, 2016; Grimmer & Stewart, 2013).

En consecuencia, el estudio combina desarrollo metodológico aplicado, procesamiento computacional de datos textuales e inferencia estadística, orientado a la generación de evidencia empírica reproducible sobre patrones discursivos en medios digitales y potencialmente transferible a otros contextos de análisis discursivo.

7.5. Fases del estudio

El presente estudio se estructura metodológicamente bajo el estándar internacional CRISP-DM (Cross Industry Standard Process for Data Mining), modelo ampliamente reconocido en ciencia de datos por su carácter iterativo, flexible y reproducible (Shearer, 2000). En esta investigación, CRISP-DM se adopta como marco operativo para el proceso de análisis de datos, complementario al diseño metodológico cuantitativo previamente descrito.

Figura 4. Metodología aplicada CRISP-DM

Nota. Elaboración propia.

Aunque CRISP-DM fue originalmente concebido para entornos industriales, su lógica secuencial y cíclica resulta adecuada para investigaciones académicas basadas en análisis de datos textuales, al permitir articular formulación del problema, preparación del corpus, modelado computacional y validación bajo un esquema reproducible.

En el contexto de esta investigación, las seis fases del modelo se adaptan al análisis del discurso mediático digital sobre corrupción en salud en Colombia (2022–2023), manteniendo coherencia con las hipótesis y variables previamente definidas.

Comprensión del problema de investigación (Entendimiento del negocio)

Corresponde a la delimitación del fenómeno de estudio, la formulación de la pregunta de investigación y la definición de objetivos e hipótesis. Se establece como propósito central identificar estructuras temáticas consistentes y diferencias estadísticamente significativas en el tono emocional del discurso mediático digital sobre corrupción en salud.

Comprensión de los datos (Recolección del corpus)

Se realiza la construcción del corpus noticioso digital mediante técnicas de extracción automatizada de contenidos, consulta de sitemaps y acceso a fuentes abiertas. Se incluyeron únicamente medios cuya política de acceso público (robots.txt y mecanismos alternativos como sitemap o RSS) no prohíbe de forma expresa la extracción automatizada con fines académicos.

En caso de restricción explícita a la minería de texto o a agentes automatizados, el medio fue excluido.

El corpus final se conforma por 518 documentos publicados entre 2022 y 2023, seleccionados bajo intersección temática estricta SALUD ∩ CORRUPCIÓN.

Preparación de los datos (Limpieza y preprocesamiento)

Se aplican procedimientos de depuración estructural y lingüística orientados a garantizar consistencia analítica del corpus, incluyendo:

- Eliminación de boilerplate.
- Deduplicación mediante Bloom Filter.
- Normalización textual.
- Procesamiento lingüístico en español.
- Exclusión de textos residuales.

Esta fase asegura calidad de datos y trazabilidad computacional para el modelado posterior.

Modelado (Análisis computacional)

Se implementan técnicas de Procesamiento de Lenguaje Natural diferenciadas según su naturaleza:

- Modelado temático no supervisado mediante Latent Dirichlet Allocation (LDA), validado con coherencia semántica (C_v) métricas de coherencia y diversidad temática.
- Análisis supervisado de sentimiento basado en representaciones contextuales del lenguaje.
- Extracción de entidades mediante reconocimiento automático de entidades nombradas (NER).

- Cálculo de métricas léxicas (TF y TF-IDF).

Estas técnicas permiten transformar estructuras discursivas en variables cuantificables para análisis estadístico.

Evaluación (Validación técnica y contrastación de hipótesis)

La fase de evaluación se desarrolla en varios niveles complementarios orientados a verificar la solidez metodológica y la coherencia interpretativa de los resultados obtenidos.

- Validación interna del modelo temático mediante métricas de coherencia semántica, estabilidad y capacidad de generalización, así como triangulación con modelos alternativos no supervisados.
- Análisis del tono emocional basado en modelos supervisados y recursos léxicos, evaluando la consistencia de las distribuciones entre medios y ejes temáticos.
- Análisis cruzado de variables discursivas (temas, sentimiento, marcos narrativos y actores) con el fin de identificar patrones estructurales del discurso mediático.
- Verificación de coherencia entre los resultados computacionales y los supuestos teóricos del estudio, mediante interpretación sistemática de los outputs analíticos.

Adicionalmente, se integra un esquema de triangulación metodológica que combina evidencias provenientes de distintos procedimientos analíticos, reduciendo la dependencia de un único modelo y fortaleciendo la validez de las conclusiones.

Despliegue (Interpretación y comunicación de resultados)

Finalmente, los patrones discursivos identificados son interpretados en función de las hipótesis planteadas, considerando variaciones temáticas, diferencias en tono emocional y dinámicas temporales. Los resultados se sistematizan en visualizaciones, tablas comparativas y análisis interpretativo, con el propósito de aportar evidencia empírica replicable para el estudio del discurso mediático digital en contextos institucionales.

Esta fase articula la transición hacia el trabajo de campo y el análisis de resultados, garantizando coherencia entre el proceso computacional descrito y la interpretación sustantiva del fenómeno estudiado.

7.6. Muestra

La muestra del estudio se compone de dos niveles analíticos diferenciados: (1) el conjunto de medios digitales seleccionados como marco de fuentes informativas y (2) el corpus documental final utilizado como unidad empírica de análisis.

Marco de fuentes informativas

Se seleccionaron diez medios de comunicación digitales mediante un muestreo no probabilístico intencional por criterios (purposive sampling), con el propósito de garantizar diversidad editorial, cobertura territorial y presencia consolidada en el ecosistema informativo colombiano. La selección se estructuró en dos categorías balanceadas:

- Cinco medios de alcance nacional.
- Cinco medios de alcance regional.

La elección no tiene pretensión de representatividad estadística poblacional, sino de representatividad analítica, orientada a capturar variabilidad discursiva entre distintos niveles de cobertura territorial. En este contexto, la representatividad analítica se entiende como la capacidad de observar patrones discursivos relevantes dentro del marco seleccionado, sin inferencia hacia el universo total de medios de comunicación.

Asimismo, únicamente se incluyeron medios cuya política de acceso público (robots.txt y mecanismos como sitemap o RSS) no prohíbe de forma expresa la extracción automatizada con fines académicos, garantizando cumplimiento ético, legal y técnico en el proceso de recolección.

Tabla 5. Medios seleccionados

Medio	Cobertura	Ciudad	Sitio Web
El Espectador	Nacional	–	https://www.elespectador.com/
Infobae Colombia	Nacional	–	https://www.infobae.com/colombia/
Cambio	Nacional	–	https://cambiocolombia.com/
Cuestión Pública	Nacional	–	https://cuestionpublica.com/
Pulzo	Nacional	–	https://www.pulzo.com/
El Heraldo	Regional	Barranquilla	https://www.elheraldo.co/
Las 2 orillas	Regional	Bogotá	https://www.las2orillas.co/
El País de Cali	Regional	Cali	https://www.elpais.com.co/

Revista Metro	Regional	Cartagena	https://revistametro.co/
El Colombiano	Regional	Medellín	https://www.elcolombiano.com/

Nota. La tabla presenta los diez medios digitales seleccionados (5 nacionales y 5 regionales) que conforman el marco de fuentes del estudio.

Corpus documental final

A partir de los diez medios seleccionados, se estableció el alcance de la recolección del corpus periodístico publicado entre 2022 y 2023 y filtrados bajo una intersección temática estricta SALUD ∩ CORRUPCIÓN.

La conformación del corpus se realizó mediante un pipeline secuencial de procesamiento que incorporó filtrado automatizado inicial, filtrado temático estricto, limpieza estructural, eliminación de boilerplate y deduplicación, hasta consolidar el conjunto analítico definitivo.

Los documentos fueron sometidos a:

- Limpieza estructural.
- Eliminación de boilerplate.
- Deduplicación probabilística mediante Bloom Filter.
- Validación temática estricta.

La unidad de análisis corresponde al documento periodístico individual, mientras que el medio digital opera como nivel de agregación secundaria para análisis comparativo entre categorías.

En consecuencia, la muestra no busca inferencia estadística hacia el universo total de medios colombianos, sino caracterizar patrones discursivos dentro del marco analítico delimitado por los medios de los cuales fue obtenido el corpus documental.

7.6.1. Criterios de exclusión de medios digitales

En la depuración del universo de medios se aplicaron criterios de exclusión orientados a garantizar que la muestra estuviera conformada únicamente por portales con producción

periodística verificable, accesibilidad técnica y pertinencia temática para los objetivos de la investigación. Estos criterios se definieron en coherencia con el enfoque de análisis de discurso textual mediante técnicas de Procesamiento de Lenguaje Natural.

Se excluyeron:

- Blogs o sitios personales sin respaldo editorial institucional, sin trayectoria verificable o sin actualización regular.
- Portales sin información verificable sobre autoría, origen institucional o mecanismos de validación editorial, lo que impide contrastar la autenticidad de sus publicaciones.
- Medios con línea temática no alineada con el objeto de estudio, tales como aquellos dedicados exclusivamente a entretenimiento, farándula, deportes o contenidos publicitarios, que no abordan de manera sistemática asuntos de interés público.
- Agregadores automatizados o portales sin producción periodística propia, que replican contenidos de otras fuentes sin generar material original.
- Portales inactivos o con actividad editorial discontinua durante el periodo 2022–2023, verificados mediante revisión directa de publicaciones disponibles en línea.
- Medios cuya producción se limita exclusivamente a formatos audiovisuales (televisión o radio) sin contenido textual accesible en formato digital, al no ser compatibles con el análisis automatizado de discurso escrito.
- Sitios con restricciones técnicas o legales que impiden la recolección automatizada de datos (por ejemplo, bloqueos explícitos en robots.txt o ausencia de mecanismos alternativos de acceso como sitemap o RSS).

En conjunto, estos criterios garantizan la calidad, pertinencia temática, accesibilidad técnica y comparabilidad del corpus, condiciones necesarias para la aplicación reproducible de métodos computacionales de análisis del discurso.

7.6.2. Reglas de selección de medios digitales

La selección de los medios de comunicación digitales que conforman el corpus siguió un conjunto de reglas metodológicas orientadas a garantizar representatividad analítica, diversidad editorial y viabilidad técnica. Estas reglas combinan indicadores cuantitativos de presencia digital y cobertura territorial con criterios cualitativos relacionados con relevancia periodística, independencia editorial y accesibilidad de los contenidos para su procesamiento mediante técnicas de Procesamiento de Lenguaje Natural (PLN).

En consecuencia, la selección de medios no se concibe como un paso logístico, sino como una decisión analítica que condiciona la diversidad discursiva observada, la comparabilidad entre coberturas nacionales y regionales y la trazabilidad del corpus. El estudio incorpora criterios explícitos de representatividad editorial y accesibilidad técnica documentada, con el fin de garantizar que el conjunto de fuentes responda al objeto de investigación y sea procesable bajo estándares de reproducibilidad.

El procedimiento se estructuró en dos niveles: medios de alcance nacional y regionales.

a. Selección de medios nacionales

La selección de medios nacionales se orientó a construir un corpus equilibrado entre organizaciones periodísticas consolidadas y plataformas nativas digitales, priorizando aquellas con amplia visibilidad pública, trayectoria editorial y cobertura informativa general.

El objetivo fue asegurar la presencia de fuentes influyentes en la esfera pública nacional, con diversidad de enfoques editoriales y producción sostenida de contenidos sobre asuntos de interés público.

Se establecieron las siguientes reglas de inclusión:

Regla 1. Presencia digital significativa.

Se priorizaron medios con alta visibilidad en el ecosistema informativo digital, estimada mediante indicadores de tráfico web y autoridad de dominio disponibles públicamente, por ejemplo, herramientas de analítica web como Ubersuggest (Neil Patel Digital, 2025),

utilizados únicamente como referencia comparativa y no como criterio exclusivo de selección.

Regla 2. Diversidad editorial e investigativa.

Se incluyeron medios con diferentes perfiles editoriales, incorporando al menos dos organizaciones dedicadas al periodismo investigativo independiente, con el fin de ampliar la diversidad de enfoques discursivos.

Regla 3. Inclusión de medios nativos digitales.

Se consideraron plataformas nacidas en entornos digitales que han alcanzado penetración significativa en audiencias nacionales, reflejando transformaciones recientes en el consumo informativo.

La aplicación conjunta de estas reglas permitió seleccionar cinco medios nacionales con alta visibilidad pública y diversidad editorial.

b. Selección de medios regionales (5 ciudades principales)

Para la selección territorial se utilizaron las proyecciones poblacionales oficiales del Departamento Administrativo Nacional de Estadística (DANE), con el propósito de identificar los principales centros urbanos del país y maximizar la cobertura territorial del estudio.

El análisis permitió identificar las cinco ciudades con mayor población durante el periodo observado: Bogotá, Medellín, Cali, Barranquilla y Cartagena. En conjunto, estas ciudades concentran cerca del 29% de la población nacional proyectada, lo que asegura una muestra territorial con amplio alcance informativo y relevancia social (Departamento Administrativo Nacional de Estadística - DANE, 2023).

Tabla 6. Selección de territorios locales considerados en el corpus

Ciudad	2022	2023
Bogotá	7.849.206	7.883.928
Medellín	2.514.709	2.518.480
Cali	2.283.342	2.280.485
Barranquilla	1.274.895	1.277.216
Cartagena	1.006.609	1.008.354
Total 5 ciudades	**14.928.761**	**14.968.463**

Población nacional	**51.643.565**	**52.117.067**
Representatividad	**28,9%**	**28,7%**

Nota. Elaborada con base en DANE, Proyecciones de Población y Estudios Demográficos (PPED), actualización consultada al 30 de julio de 2025.

Reglas de selección de medios regionales

A partir de estas ciudades se aplicaron los siguientes criterios:

Regla 1. Representatividad informativa local.

Para cada ciudad se identificaron medios digitales con trayectoria editorial, producción periodística regular y reconocimiento en el ámbito regional.

Regla 2. Cobertura de asuntos públicos.

Los medios debían publicar de manera sistemática contenidos políticos, sociales o institucionales, incluyendo información relacionada con el sector salud.

Regla 3. Exclusión de medios de baja incidencia.

Se descartaron medios universitarios o comunitarios con circulación limitada, salvo aquellos con reconocimiento institucional o alcance regional significativo.

Regla 4. Compatibilidad con análisis textual.

Se excluyeron portales dependientes exclusivamente de formatos audiovisuales, manteniendo únicamente aquellos con producción escrita disponible en formato digital.

Regla 5. Evitar duplicidad con medios nacionales.

Se impidió la repetición de fuentes entre los niveles nacional y regional.

Tras la aplicación de estos criterios se seleccionó un medio por cada ciudad, estableciendo de cinco medios regionales comparables en términos de cobertura territorial.

Este procedimiento permitió construir un conjunto de fuentes informativas equilibrado entre cobertura nacional y regional, con diversidad editorial y accesibilidad técnica suficiente para la construcción de un corpus textual homogéneo y reproducible.

7.6.3. Definición del universo potencial de medios digitales

Se elaboró un listado preliminar de 25 medios digitales colombianos activos durante el periodo 2022–2023 a partir de una estrategia de triangulación de fuentes y herramientas. Inicialmente, se consultaron registros de la Asociación Colombiana de Medios de Información (2023) y el ranking de medios publicado por Revista P&M (Revista P&M, 2023), los cuales ofrecen un panorama institucional y de posicionamiento digital en el país.

De manera complementaria, se realizó una exploración de fuentes digitales mediante búsquedas sistemáticas y verificación manual de actividad editorial, con el fin de identificar medios regionales con producción periodística continua durante 2022–2023.

Finalmente, se realizó una validación directa de la existencia y actividad digital de cada medio en internet, verificando su actualización durante el periodo analizado y descartando aquellos portales inactivos o sin continuidad periodística comprobable. Con base en este proceso, se consolidó el universo de medios que posteriormente fue depurado para la selección final empleada en el análisis.

Tabla 7. Preselección de medios digitales

Medio	Cobertura	Ciudad	Sitio Web
El Espectador	Nacional	–	https://www.elespectador.com/
Semana	Nacional	–	https://www.semana.com/
El Tiempo	Nacional	–	https://www.eltiempo.com/
Infobae Colombia	Nacional	–	https://www.infobae.com/colombia/
La Silla Vacía	Nacional	–	https://www.lasillavacia.com/
La República	Nacional	–	https://www.larepublica.co/
Cambio	Nacional	–	https://cambiocolombia.com/
Cuestión Pública	Nacional	–	https://cuestionpublica.com/
Pulzo	Nacional	–	https://www.pulzo.com/
Vorágine	Nacional	–	https://voragine.co/
El Heraldo	Local	Barranquilla	https://www.elheraldo.co/
Zona Cero	Local	Barranquilla	https://zonacero.com/
Diario La Libertad	Local	Barranquilla	https://diariolalibertad.com
Las2orillas	Local	Bogotá	https://www.las2orillas.co/
KienyKe	Local	Bogotá	https://www.kienyke.com/
El Nuevo Siglo	Local	Bogotá	https://www.elnuevosiglo.com.co/

El País de Cali	Local	Cali	https://www.elpais.com.co/
Diario Occidente	Local	Cali	https://occidente.co/
Noti90 Minutos	Local	Cali	https://90minutos.co/
El Universal	Local	Cartagena	https://www.eluniversal.com.co/
Revista Metro	Local	Cartagena	https://revistametro.co/
El Bolivarense	Local	Cartagena	https://bolivarense.com/
El Colombiano	Local	Medellín	https://www.elcolombiano.com/
Minuto30	Local	Medellín	https://www.minuto30.com/
El Mundo	Local	Medellín	http://www.elmundo.com/

Nota. La tabla presenta los 25 medios digitales preseleccionados (10 nacionales y 15 regionales) que conforman el universo potencial de análisis previo a la selección final.

Este universo inicial no pretende exhaustividad absoluta, sino establecer una cobertura amplia del ecosistema informativo digital colombiano con capacidad de producción periodística durante el periodo analizado.

7.6.4. Fuente de información y variables de selección

Las métricas de tráfico orgánico, autoridad de dominio, palabras clave orgánicas y backlinks se obtuvieron a partir de consultas realizadas en la herramienta pública Neil Patel – Ubersuggest (*Website Traffic Checker*), que permite estimar el rendimiento digital de los dominios y su posicionamiento en buscadores (Neil Patel Digital, 2025).

Tabla 8. Variables de selección de medios digitales

Variable	Descripción	Propósito en la selección del medio
Tráfico orgánico mensual estimado	Estimación del volumen de visitas provenientes de motores de búsqueda.	Identificar el alcance real del medio y su visibilidad en la web.
Palabras clave orgánicas	Número de términos por los cuales el dominio se posiciona de manera orgánica.	Medir la amplitud del posicionamiento y robustez SEO del medio.
Autoridad del dominio (según métricas SEO reportadas por Ubersuggest)	Indicador de autoridad del dominio como proxy de posicionamiento y visibilidad digital (métrica SEO).	Evaluar la reputación digital y la solidez estructural del medio.
Número de backlinks	Cantidad de enlaces entrantes desde otros sitios web.	Medir reputación, difusión y relevancia en el ecosistema digital.
Palabras clave patrocinadas (SEM)	Cantidad de términos posicionados mediante pago (SEM).	Distinguir tráfico genuino/orgánico del tráfico comprado.

Distribución geográfica del tráfico por país	Lista de las URLs más visitadas y su distribución geográfica.	Validar si el tráfico es mayoritariamente nacional (criterio clave del estudio).
Principales páginas visitadas (SEO)	URLs con mayor tráfico por búsquedas orgánicas.	Confirmar que el medio publica contenido relevante para “corrupción en salud”.
Principales palabras clave SEO	Términos más frecuentes por los cuales llega tráfico al sitio.	Analizar la pertinencia temática y el comportamiento del posicionamiento.

Nota. Elaborada con base en métricas de rendimiento digital obtenidas mediante la herramienta Ubersuggest (Neil Patel Digital, 2025).

Estas métricas se emplean como indicadores de alcance digital, no como validación de calidad periodística, se utilizaron de forma comparativa entre dominios y no como valores absolutos, dado que las estimaciones de tráfico web varían según la metodología de cada herramienta de analítica digital.

Se obtuvieron consultando el dominio oficial de cada medio en Neil Patel – Ubersuggest, registrando las métricas reportadas para Colombia para el 18 de agosto de 2025: tráfico orgánico mensual, palabras clave orgánicas, autoridad del dominio, backlinks, palabras clave patrocinadas, principales páginas visitadas y principales palabras clave. Se documenta la fecha de consulta para asegurar la reproducibilidad del procedimiento.

7.6.5. Alcance general de la muestra

La investigación parte de una muestra preliminar de 25 medios digitales (10 nacionales y 15 regionales, tres por ciudad). Sobre este conjunto se aplica posteriormente una matriz de evaluación y un filtro de viabilidad técnica y legal, incluyendo la revisión de políticas de acceso automatizado (robots.txt), con el fin de definir el conjunto de fuentes que conforman el corpus.

El procedimiento corresponde a un muestreo no probabilístico intencionado estructurado, orientado a maximizar diversidad discursiva y viabilidad de procesamiento.

Tabla 9. Composición general de la muestra y variables de evaluación

Categoría de medio	Cantidad	Criterio de inclusión	Descripción
Nacionales	10	Alta cobertura informativa y/o periodismo investigativo.	Medios digitales con amplio alcance, autoridad de dominio y trayectoria en temas de corrupción y salud.

Locales / Regionales	15 (3 por ciudad)	Representatividad territorial.	Medios reconocidos en las cinco ciudades principales: Bogotá, Medellín, Cali, Barranquilla y Cartagena.
Muestra inicial	25 medios digitales	Equilibrio entre alcance nacional y regional.	Selección orientada a garantizar diversidad editorial y viabilidad técnica para el análisis automatizado.

Nota. La tabla presenta la estructura inicial de la muestra seleccionada para el estudio, basada en criterios de cobertura editorial, relevancia investigativa, representatividad territorial y viabilidad técnica para el procesamiento automatizado (robots.txt). Las variables usadas en esa evaluación se presentan en la Tabla 10.

Tabla 10. Variables consideradas en la matriz de evaluación de medios digitales

Variable	Escala	Descripción / Propósito
Tráfico orgánico (TRA)	1–5	Mide el volumen de visitas mensuales provenientes de buscadores.
Autoridad del dominio (según métricas SEO)	1–5	Evalúa la reputación y confiabilidad del sitio según su índice SEO.
Palabras clave orgánicas (KW)	1–5	Refleja la amplitud del posicionamiento del medio en buscadores.
Backlinks (BL)	1–5	Indica la cantidad de enlaces externos que refieren al medio.
Relevancia investigativa (INV)	1–5	Evaluación cualitativa experta de la especialización del medio en periodismo investigativo y cobertura de corrupción/salud.
Puntaje total	0–5 (normalizado)	Resultado de la fórmula de evaluación aplicada a todas las variables.
Nivel de acceso (robots.txt)	Categórico: Permite / Parcial / Restringido	Determina la viabilidad técnica y legal para la recolección automatizada, conforme a estándares de recolección responsable de datos.
Selección final	Incluido / Excluido	Define la participación del medio en el corpus final.

Nota. Las variables presentadas corresponden a los criterios utilizados para evaluar los 25 medios digitales incluidos en la matriz de selección. Las escalas y descripciones se basan en métricas de posicionamiento web obtenidas mediante Ubersuggest (Neil Patel Digital, 2025) y en criterios de relevancia definidos en el diseño metodológico del estudio.

7.6.6. Variables y pesos

Se evaluó cada medio mediante cinco variables principales de naturaleza cuantitativa y cualitativa. La puntuación total corresponde a una suma ponderada en el rango 0–5.

Fórmula general:

Total = 0.45 * INV + 0.20 * TRA + 0.15 * BL + 0.10 * DA + 0.10 * KW

Donde

- INV = Relevancia investigativa (alineación con corrupción / salud, investigación, profundidad) – peso 45%
- TRA = Tráfico orgánico mensual (visitas desde buscadores) – peso 20%
- BL = Backlinks (enlaces entrantes) – peso 15%
- DA = Indicador de autoridad del dominio (escala SEO 0–100) – peso 10%
- KW = Palabras clave orgánicas posicionadas – peso 10%

La ponderación se definió ex ante para priorizar la pertinencia sustantiva del medio respecto del fenómeno analizado (INV) y, en segundo plano, su alcance y visibilidad digital (TRA, BL, DA, KW), evitando que la selección quedara dominada únicamente por métricas SEO.

Se asignó mayor peso a INV (45%) porque el objetivo del estudio busca análisis de discurso en corrupción en salud y se requiere priorizar medios con investigación y profundidad, mientras que las métricas SEO (TRA, BL, DA, KW) aportan una medida complementaria de alcance y posicionamiento digital.

Todas las variables fueron normalizadas a una escala ordinal de 1 a 5 antes del cálculo ponderado.

A modo ilustrativo, si un medio presenta INV = 5, TRA = 4, BL = 3, DA = 4 y KW = 4:

Total = 0.45·5 + 0.20·4 + 0.15·3 + 0.10·4 + 0.10·4

Total = 2.25 + 0.80 + 0.45 + 0.40 + 0.40 = **4.30 / 5.00**

Variable cualitativa clave: Relevancia investigativa (INV)

La evaluación se realizó mediante revisión sistemática de contenidos publicados por cada medio durante el periodo de referencia aplicando los criterios operativos definidos en la Tabla 11.

Definición (1–5) — se evalúa el alineamiento del medio con periodismo investigativo y cobertura de corrupción/sector salud:

Tabla 11. Criterios de evaluación para la variable “Relevancia investigativa (INV)”

Puntaje (1–5)	Descripción operativa	Criterios específicos
5	Medio con enfoque investigativo nativo (anticorrupción, transparencia, investigación profunda).	• Series de reportajes. • Especiales sobre corrupción o salud. • Bases de datos, líneas temáticas o proyectos permanentes. • Alta recurrencia en investigaciones estructurales.
4	Medio generalista con unidad de investigación consolidada.	• Publicaciones regulares de investigaciones largas. • Dossiers temáticos. • Verificación y cruce documental. • Historial de investigaciones relevantes en corrupción/salud.
3	Medio generalista con cobertura frecuente del tema, pero sin unidad investigativa formal.	• Noticias procesadas + análisis. • Reportajes esporádicos. • Secciones temáticas estables, pero no investigativas.
2	Medio generalista con cobertura ocasional de corrupción o salud.	• Noticias de coyuntura. • Cobertura reactiva. • Enfoque informativo más que investigativo.
1	Medio agregador o de noticias breves; mínimo trabajo investigativo.	• Reproducción de boletines. • Cobertura superficial. • Ausencia total de investigación propia.

Nota. La tabla presenta la definición operacional de la variable cualitativa “Relevancia investigativa (INV)”, utilizada para evaluar el perfil editorial de los 25 medios incluidos en la muestra. Los criterios se construyeron a partir de la trayectoria investigativa, presencia de unidades de investigación, producción de reportajes y frecuencia de cobertura en temas de corrupción y salud.

Reglas de escalamiento (1–5) para las variables cuantitativas

Estas reglas convierten el valor bruto (de Ubersuggest / Neil Patel) a escala 1–5.

Tabla 12. Umbrales de evaluación para métricas SEO (TRA, DA, KW y BL)

Métrica	Puntaje	Umbral	Descripción / Interpretación
Tráfico orgánico mensual (TRA)	5	> 3.000.000 visitas / mes	Alcance digital muy alto; medios líderes en audiencia.
	4	1.000.000 – 3.000.000	Alto posicionamiento y visibilidad en búsquedas.

	3	500.000 – 999.999	Alcance medio-alto; fuerte presencia en buscadores.
	2	100.000 – 499.999	Alcance medio; tráfico estable pero limitado.
	1	< 100.000	Bajo alcance orgánico; visibilidad limitada.
Indicador de autoridad del dominio (según métricas SEO)	5	> 85	Máxima confiabilidad y reputación SEO
	4	75 – 85	Dominio robusto, estable y ampliamente referenciado.
	3	60 – 74	Autoridad sólida; buen posicionamiento en el ecosistema digital.
	2	45 – 59	Reputación moderada; fortalecimiento en curso.
	1	< 45	Dominio débil; limitada reputación digital.
Palabras clave orgánicas (KW)	5	> 1.000.000 keywords	Posicionamiento masivo en buscadores
	4	500.000 – 1.000.000	Amplio espectro de posicionamiento orgánico.
	3	100.000 – 499.999	Visibilidad media estable.
	2	10.000 – 99.999	Cobertura temática limitada.
	1	< 10.000	Bajo posicionamiento SEO.
Backlinks (BL)	5	> 10.000.000 enlaces entrantes	Máxima reputación digital y alta referenciación externa.
	4	5.000.000 – 10.000.000	Fuerte presencia en el ecosistema web.
	3	1.000.000 – 4.999.999	Referenciación media-alta.
	2	100.000 – 999.999	Número moderado de enlaces entrantes.
	1	< 100.000	Poca referenciación desde otros dominios.

Nota. Los umbrales presentados corresponden a la escala utilizada para evaluar métricas SEO de los 25 medios incluidos en la muestra. Los valores derivan de estimaciones obtenidas mediante Ubersuggest (Neil Patel Digital, 2025).

Variable complementaria: Accesibilidad técnica y legal

Además de las variables de desempeño y relevancia, se incorporó un criterio complementario de accesibilidad técnica basado en la revisión de robots.txt y, cuando aplica, en restricciones públicas de acceso (por ejemplo, sitemaps o RSS). Este criterio se usa como señal operativa para minimizar fricciones técnicas y reducir el riesgo de recolección indebida en sitios que declaran restricciones a agentes automatizados. En consecuencia, el estudio prioriza fuentes con acceso abierto o semiautomatizable, y documenta estas condiciones para efectos de trazabilidad y reproducibilidad del proceso.

Solo se consideraron aptos los medios cuya política de acceso público (robots.txt y mecanismos alternativos como sitemap/RSS) no declara restricciones al acceso automatizado con fines académicos, o que ofrecen mecanismos abiertos de acceso y preservación que habilitan la recolección sin afectar su infraestructura.

Este procedimiento asegura la integridad del corpus, el cumplimiento de los lineamientos éticos para investigación en internet (Association of Internet Researchers, 2019) y la reproducibilidad científica dentro de los márgenes legales establecidos por los medios. El análisis de los archivos robots.txt se aplicó como una verificación obligatoria para identificar posibles restricciones y validar que la recolección de datos se limite a fuentes legítimas.

En términos de accesibilidad técnica, los portales se clasificaron según las condiciones de sus archivos robots.txt:

- **Permite:** acceso mediante scraping o sitemap público.
- **Parcial:** restringe algunos agentes automatizados específicos, pero permite acceso mediante sitemap o RSS.
- **Restringido:** prohíbe expresamente la minería de texto o el uso de bots; por tanto, excluido del corpus.

Esta clasificación refuerza la transparencia, trazabilidad y legalidad del proceso de recolección, asegurando la conformidad del proyecto con los principios de ética digital y uso responsable de información pública. El criterio robots.txt se aplica como filtro obligatorio: si el medio restringe explícitamente la recolección automatizada, se excluye, independientemente del puntaje total.

Las métricas SEO utilizadas corresponden a mediciones puntuales realizadas en 2025 y se emplean únicamente como *proxy* de presencia digital y visibilidad relativa del dominio, no como estimación histórica del desempeño durante 2022–2023.

7.6.7. Tamaño final de la muestra

El tamaño final de la muestra estuvo conformado por diez medios de comunicación digitales, seleccionados bajo un criterio de equilibrio entre alcance nacional (cinco medios) y cobertura regional (cinco medios ubicados en Bogotá, Medellín, Cali, Barranquilla y Cartagena). Esta distribución garantiza cobertura territorial analítica y diversidad editorial, incorporando medios investigativos y plataformas digitales independientes con líneas editoriales diferenciadas.

La definición de esta muestra se fundamenta en tres criterios centrales:

- Diversidad de perspectivas: medios nacionales y regionales con distintas orientaciones editoriales, lo que garantiza heterogeneidad discursiva y comparabilidad narrativa.
- Cobertura temática suficiente: volumen informativo suficiente para capturar variaciones en el tratamiento periodístico del fenómeno dentro del corpus analizado.
- Viabilidad metodológica: mantenimiento de un corpus de textos, adecuado para el procesamiento automatizado mediante técnicas de PLN.

Este muestreo equilibra representatividad, diversidad y factibilidad técnica, asegurando la coherencia entre los objetivos de investigación y la capacidad analítica del estudio.

En términos operacionales, el filtro de accesibilidad técnica y legal (robots.txt y mecanismos alternativos como sitemap o RSS) se aplicó como condición habilitante previa a la evaluación por puntaje; posteriormente, se aplicó la ponderación definida para priorizar medios con mayor pertinencia investigativa y presencia digital.

Reglas de selección y desempate

Para pasar de la muestra inicial (25 medios) a la muestra final (10 medios), se siguieron los siguientes pasos: (1) se calculó el puntaje total para cada medio con la fórmula y escalas definidas; (2) se aplicó el filtro de acceso robots.txt; (3) se seleccionaron los 5 nacionales con mayor puntaje y los 5 regionales (uno por ciudad) usando las reglas de la Tabla 13.

Tabla 13. Reglas de selección y criterios de desempate para medios nacionales y regionales

Categoría	Criterio principal	Criterios complementarios y desempate
Medios nacionales (5)	Seleccionados según el mayor puntaje total y nivel de accesibilidad técnica “Permite” o “Parcial”.	Deben incluir al menos dos medios con alta relevancia investigativa (INV ≥ 4).
Medios regionales (5)	Se elige uno por ciudad (Bogotá, Medellín, Barranquilla, Cali y Cartagena), priorizando el mayor puntaje total y acceso “Permite” o “Parcial”.	En caso de empate, se prioriza primero la relevancia investigativa (INV) y luego el tráfico orgánico (TRA).

Nota. La tabla resume los criterios aplicados para la selección final de medios nacionales y regionales, integrando puntaje total, nivel de acceso técnico (robots.txt) y las reglas de desempate basadas en relevancia investigativa (INV) y tráfico orgánico (TRA). Para evitar ambigüedades, en esta investigación solo se consideran aptos los medios con robots.txt en nivel ‘Permite’ y ‘Parcial’. El nivel ‘Restringido’ se excluye.

Además del puntaje total, como condición adicional, se exige que la selección de los 5 medios nacionales incluya al menos dos con alta relevancia investigativa (INV ≥ 4). Este criterio busca asegurar que el corpus no esté compuesto únicamente por medios con alto alcance digital, sino que incorpore medios con capacidad de investigación y profundidad temática, fundamentales para el análisis de discurso sobre corrupción en el sector salud. En caso de que el Top-5 nacional por puntaje no cumpla esta condición, se reemplaza el medio de menor puntaje por el siguiente medio disponible que cumpla INV ≥ 4 y tenga acceso permitido o parcial, hasta satisfacer la regla.

En síntesis, la muestra del estudio se definió a partir de un universo inicial de 25 medios digitales, depurado mediante criterios editoriales, métricas de presencia digital y verificación obligatoria de accesibilidad técnica y legal (robots.txt). Este proceso condujo a la selección final de 10 medios (5 nacionales y 5 regionales) y a la conformación de un corpus analítico de documentos publicados entre 2022 y 2023, filtrados bajo intersección temática estricta SALUD ∩ CORRUPCIÓN y depurados mediante limpieza estructural y deduplicación, garantizando trazabilidad, reproducibilidad y comparabilidad analítica entre fuentes.

Tabla 14. Matriz de evaluación de medios digitales

Medio	Cobertura	Ciudad	Tráfico (1–5)	DA (1–5)	Keywords (1–5)	Backlinks (1–5)	Investigación	Puntaje total	Nivel de acceso	Selección
El Espectador	Nacional	–	5	5	5	5	4	4,55	Permite	☑
Semana	Nacional	–	5	5	5	4	4	4,40	Parcial	↔
El Tiempo	Nacional	–	5	5	5	5	3	4,10	Restringido	☒
Infobae Colombia	Nacional	–	5	5	5	5	3	4,10	Permite	☑
La Silla Vacía	Nacional	–	3	3	3	3	5	3,90	Parcial	↔
La República	Nacional	–	5	4	5	4	3	3,85	Parcial	↔
Cambio	Nacional	–	4	3	3	3	4	3,65	Parcial	↔
Cuestión Pública	Nacional	–	2	2	3	3	5	3,60	Permite	☑
Pulzo	Nacional	–	5	5	5	4	2	3,50	Parcial	↔
Vorágine	Nacional	–	2	2	3	2	5	3,45	Permite	☑
El Heraldo	Local	Barranquilla	5	5	5	5	3	4,10	Permite	☑
Zona Cero	Local	Barranquilla	3	3	3	3	3	3,00	Permite	☑
Diario La Libertad	Local	Barranquilla	2	2	2	2	2	2,00	Restringido	☒
Las2orillas	Local	Bogotá	4	4	4	4	4	4,00	Permite	☑
KienyKe	Local	Bogotá	4	4	4	4	3	3,55	Parcial	↔
El Nuevo Siglo	Local	Bogotá	3	3	3	3	3	3,00	Permite	☑
El País de Cali	Local	Cali	5	5	5	5	3	4,10	Permite	☑
Diario Occidente	Local	Cali	3	3	3	3	3	3,00	Permite	☑
Noti90 Minutos	Local	Cali	3	3	3	3	2	2,55	Parcial	↔
El Universal	Local	Cartagena	4	4	4	4	3	3,55	Parcial	↔
Revista Metro	Local	Cartagena	1	1	1	1	3	2,10	Permite	☑
El Bolivarense	Local	Cartagena	1	1	1	1	2	1,65	Permite	☑
El Colombiano	Local	Medellín	5	5	5	5	4	4,55	Permite	☑
Minuto30	Local	Medellín	4	4	4	4	2	3,20	Permite	☑
El Mundo	Local	Medellín	3	3	3	3	3	3,00	Restringido	☒

Nota. Elaborada con base en mediciones puntuales realizadas en 2025 mediante Neil Patel – Ubersuggest (Website Traffic Checker) y cálculos derivados de la metodología de ponderación definida en el estudio. La tabla consolida las métricas de tráfico orgánico, autoridad de dominio, palabras clave, backlinks y relevancia investigativa y nivel de acceso de los medios digitales seleccionados.

7.7. Instrumento de medición y procesamiento

El instrumento diseñado para esta investigación corresponde a un instrumento computacional de medición indirecta desarrollado en Python, estructurado bajo una arquitectura modular que integra dos subsistemas principales: (1) un módulo de ingeniería de datos para la recolección automatizada del corpus documental y (2) un módulo de procesamiento de lenguaje natural (PLN) para el análisis discursivo automatizado.

Este instrumento se enmarca en las fases de comprensión de los datos, preparación, modelado y evaluación del modelo CRISP-DM, garantizando trazabilidad, replicabilidad y consistencia metodológica en el tratamiento del corpus.

7.7.1. Subsistema de Recolección de Datos

Para la construcción del corpus documental se implementó una arquitectura basada en el framework Scrapy (v2.13), orientada a la extracción automatizada de contenido públicamente accesible en medios digitales cuyos archivos robots.txt no restringen la recolección automatizada con fines académicos.

El sistema opera mediante estrategias diferenciadas, adaptadas a la arquitectura tecnológica de cada medio:

Extracción mediante parsing HTML (selectores CSS o XPath): Aplicada en medios con estructura estándar de publicación digital, permitiendo aislar cuerpo de noticia, fecha y metadatos relevantes.

Consulta programática a endpoints REST públicos: En medios con arquitectura basada en WordPress, se emplearon endpoints oficiales (wp-json/wp/v2/posts) para recuperar información estructurada en formato JSON.

Recuperación histórica mediante Internet Archive (API CDX): En casos de indisponibilidad temporal del contenido en el sitio activo, se consultaron copias archivadas correspondientes al periodo 2022–2023, garantizando integridad temporal del corpus.

En todos los casos, la extracción se limitó a contenido de acceso público y conforme a las condiciones de accesibilidad técnica previamente definidas, respetando lineamientos éticos y técnicos de recolección automatizada.

7.7.2. Pipeline de filtrado temático estricto

Posterior a la extracción inicial, se implementó un pipeline de validación temática basado en la intersección lógica de conjuntos léxicos asociados a los dominios SALUD y CORRUPCIÓN, implementado de forma automatizada sobre el texto completo.

Un documento se considera válido cuando contiene al menos un término perteneciente al conjunto léxico asociado al dominio salud y, simultáneamente, al menos un término perteneciente al conjunto asociado a corrupción. Formalmente:

$$d \text{ válido} \Leftrightarrow (\exists\, k_1 \in K_{Salud} : k_1 \in d) \wedge (\exists\, k_2 \in K_{Corrupción} : k_2 \in d)$$

donde:

- $\boldsymbol{K_{Salud}}$ representa el conjunto de palabras clave asociadas al dominio sanitario
- $\boldsymbol{K_{Corrupción}}$ representa el conjunto de términos vinculados a prácticas corruptas
- **d** corresponde al contenido textual íntegro del documento periodístico

Estos conjuntos léxicos se construyeron a partir de vocabulario especializado, términos institucionales y expresiones frecuentes en la cobertura periodística del sector salud y de fenómenos de corrupción.

Este criterio de intersección estricta garantiza la coherencia temática del corpus y evita la inclusión de documentos tangenciales que mencionen solo uno de los dominios. En este contexto, el documento periodístico individual constituye la unidad mínima de análisis.

La selección del corpus final se ejecutó mediante un pipeline secuencial de procesamiento de datos que integró distintas etapas de depuración y validación. El procedimiento incluyó una preclasificación automática durante la fase de recolección, seguida de un filtrado temático estricto basado en la intersección de términos asociados al sector salud

y a fenómenos de corrupción. Posteriormente se aplicaron procesos de limpieza estructural del texto, eliminación de contenido boilerplate y deduplicación de artículos.

Este flujo de procesamiento permitió consolidar el conjunto analítico definitivo utilizado en las etapas posteriores de modelado temático, análisis de sentimiento y aproximación computacional al framing discursivo.

Tras la aplicación de los criterios de filtrado temático, limpieza textual, eliminación de contenido boilerplate y deduplicación de artículos, el corpus final analizado quedó conformado por 518 artículos periodísticos publicados entre 2022 y 2023, los cuales constituyen la base empírica para las etapas posteriores de análisis.

7.7.3. Subsistema de Análisis y Procesamiento (PLN)

Una vez consolidado el corpus documental validado, el procesamiento analítico se ejecutó mediante un flujo de procesamiento secuencial estructurado en fases interdependientes, garantizando trazabilidad desde el dato bruto hasta la generación de variables analíticas derivadas.

Unificación y exploración inicial

Los archivos estructurados resultantes del proceso de recolección fueron consolidados en una estructura tabular en formato columnar Parquet, gestionada mediante la biblioteca Pandas, con identificadores únicos por documento, permitiendo gestión eficiente de memoria y replicabilidad del flujo analítico.

En esta etapa se realizó una exploración descriptiva preliminar, incluyendo distribución temporal de publicaciones y volumen documental por medio digital, con fines de control de calidad y caracterización inicial del corpus.

Preprocesamiento textual

Se aplicaron procedimientos de normalización lingüística orientados al modelado semántico, incluyendo:

- Conversión a minúsculas.
- Eliminación de caracteres no alfabéticos.
- Tokenización.
- Filtrado de stopwords en español.
- Lematización.

Estas operaciones se implementaron mediante librerías especializadas (NLTK y spaCy), generando un corpus optimizado para modelado probabilístico, minimizando la pérdida de información semántica relevante.

Modelado de Tópicos (LDA)

Se implementó el algoritmo Latent Dirichlet Allocation (LDA) sobre el corpus preprocesado, a partir de una representación de bolsa de palabras (Bag-of-Words, BoW) y un diccionario léxico derivado del corpus.

El modelo final fue seleccionado mediante un esquema iterativo de ajuste de hiperparámetros, utilizando como criterios:

- Coherencia semántica C_v.
- Coherencia C_npmi como medida complementaria de robustez léxica.
- Diversidad temática (Topic Diversity).
- Consistencia entre ejecuciones.

Cada documento recibió una distribución probabilística sobre los tópicos generados y un tópico dominante, operacionalizado como variable categórica con probabilidad asociada.

Análisis de Sentimiento

El análisis del tono emocional se realizó mediante un modelo contextual basado en transformadores preentrenados para español (RoBERTa), implementado a través de la biblioteca pysentimiento, seleccionado por su capacidad para capturar dependencias semánticas complejas y evaluaciones implícitas propias del discurso periodístico.

Con fines de validación metodológica, se aplicaron adicionalmente dos baselines léxicos ampliamente utilizados en PLN para español:

- ML-SentiCon.
- NRC Word-Emotion Association Lexicon (EmoLex) en su versión en español.

Estos recursos se aplicaron sobre el mismo corpus base ya limpiado y filtrado y se compararon con el modelo contextual mediante métricas de correlación, acuerdo de etiquetas y consistencia de ranking, con el objetivo de evaluar la estabilidad de los resultados sin sustituir el modelo principal.

Adicionalmente, se incorporó un diagnóstico de incertidumbre predictiva y desacuerdo metodológico, orientado a evaluar la confianza del modelo contextual y su convergencia parcial con los baselines léxicos.

A diferencia del modelado temático probabilístico, los modelos basados en representaciones contextuales no emplearon lematización como insumo principal. En BERTopic se utilizó limpieza básica del texto, dado que la representación semántica se deriva de embeddings contextuales, y en el análisis de sentimiento con RoBERTa se trabajó sobre el texto limpio en su forma secuencial, preservando el contexto léxico necesario para la inferencia contextual del modelo.

Integración y cruce de variables

Los resultados del modelado temático y del análisis de sentimiento fueron integrados en un dataset enriquecido, asignando a cada documento:

- Tópico dominante.
- Distribución probabilística completa.
- Score de polaridad emocional.
- Medio de publicación.
- Fecha de publicación.

Esto permitió la construcción de tablas de contingencia y análisis agregados, incluyendo:

- Sentimiento promedio por medio.
- Distribución temática por medio.
- Evolución temporal de tópicos.
- Cruces tópico × medio.

Validación estadística e interpretación

La contrastación de la hipótesis se sustentó en la convergencia de evidencias obtenidas mediante los distintos componentes analíticos del estudio, incluyendo el modelado temático, el análisis de sentimiento y la identificación de marcos narrativos y actores. La robustez de los resultados se evaluó a partir de métricas internas de los modelos, análisis de estabilidad y triangulación metodológica entre enfoques alternativos.

En particular, la consistencia de las distribuciones del tono emocional entre medios y ejes temáticos se examinó mediante análisis descriptivos y comparativos, apoyados en modelos supervisados basados en representaciones contextuales y en recursos léxicos especializados.

Finalmente, se realizó la interpretación semántica de los tópicos identificados, asignando etiquetas conceptuales coherentes con el marco teórico del framing mediático. Este proceso se llevó a cabo mediante una interpretación guiada por la literatura y por los términos de mayor peso probabilístico en cada tópico, con el fin de garantizar correspondencia entre los resultados computacionales y el análisis discursivo.

7.7.4. Validación del componente de modelado temático del sistema computacional de análisis discursivo

La validación del componente de modelado temático del sistema computacional se realizó mediante procedimientos estrictamente cuantitativos, apoyados en métricas de

coherencia interna y estabilidad semántica ampliamente aceptadas en la literatura sobre modelado temático y Procesamiento de Lenguaje Natural (PLN) (Lau et al., 2014; Röder et al., 2015). Dado que el objetivo del estudio se centra en la identificación de estructuras temáticas latentes y patrones discursivos recurrentes, la validación se orientó a garantizar interpretabilidad semántica, separación estructural y consistencia estadística del modelo, más que exactitud predictiva, en coherencia con enfoques no supervisados aplicados a datos textuales en ciencias sociales (Grimmer & Stewart, 2013).

La coherencia temática se utiliza como proxy cuantitativo de interpretabilidad semántica de los tópicos generados.

Los resultados del modelo LDA fueron posteriormente contrastados con modelos alternativos (HDP y BERTopic) con fines de triangulación metodológica y evaluación de estabilidad temática.

La validación del modelo Latent Dirichlet Allocation (LDA) se desarrolló mediante un proceso iterativo de optimización de hiperparámetros, evaluando distintas configuraciones del número de tópicos (k) bajo un esquema controlado de filtrado léxico (no_below y no_above).

Cada configuración fue evaluada utilizando:

- Coherencia semántica C_v, seleccionada por su alta correlación con la interpretabilidad humana de los tópicos (Röder et al., 2015).
- Coherencia C_npmi como medida complementaria de robustez léxica basada en la asociación entre términos.
- Diversidad temática (Topic Diversity, TD), orientada a medir el grado de diferenciación léxica entre los términos dominantes de los tópicos generados y a detectar redundancia semántica entre ellos.

Como resultado del proceso de búsqueda, se identificó como configuración óptima el modelo con:

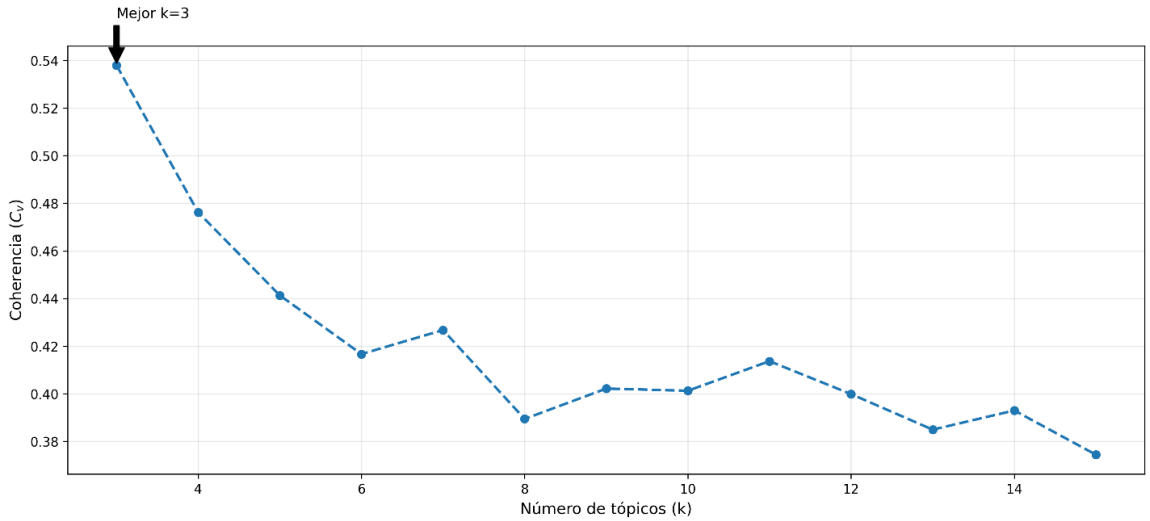
- k = 3 tópicos
- no_below = 3
- no_above = 0.4
- C_v = 0.5379
- C_npmi = 0.0256
- TD = 0.8250

El valor de coherencia obtenido corresponde al mayor valor observado dentro del rango evaluado de configuraciones, lo que respalda empíricamente la selección de k = 3 como configuración óptima.

El valor de diversidad temática (TD = 0.8250) indica un alto grado de diferenciación léxica entre los tópicos, lo que sugiere una separación semántica sustantiva del espacio discursivo sin redundancia significativa entre conjuntos de términos dominantes.

Valores cercanos a 1 indican menor redundancia léxica entre tópicos y mayor separación semántica.

Figura 5. Evolución coherencia semántica (C_v) del modelo LDA según número de tópicos (k)



Nota. La figura muestra la variación de la coherencia semántica (C_v) en función del número de tópicos evaluados. El valor máximo se alcanza en k = 3, criterio utilizado para la selección del modelo óptimo.

Asimismo, se verificó la estabilidad del modelo mediante:

- Persistencia de términos dominantes por tópico en iteraciones sucesivas.
- Conservación de artefactos fundamentales (estado del modelo LDA, diccionario léxico y matrices internas).
- Reproducibilidad del pipeline bajo condiciones controladas de ejecución, incluyendo fijación de semilla aleatoria y preservación de artefactos intermedios.

Una vez validado el modelo, se asignó a cada documento su distribución probabilística de tópicos y su tópico dominante, generando el dataset enriquecido que sustenta los análisis posteriores de framing mediático, contraste de tono emocional y evolución temporal.

Es importante señalar que la interpretación semántica de los tópicos se realizó en una fase analítica posterior con fines explicativos, y no constituyó un criterio adicional para la selección del modelo, preservando así la objetividad del proceso de validación estadística.

Adicionalmente, se incorporó un diagnóstico de ajuste y estabilidad basado en evaluación por particiones, análisis de sensibilidad al número de tópicos y remuestreo bootstrap, con el fin de identificar señales de subajuste, sobresegmentación e inestabilidad estructural en el bloque temático, para evaluar la robustez del modelo frente a variaciones de muestreo y del número de tópicos.

En conjunto, este esquema proporciona evidencia de coherencia interna, estabilidad estructural y reproducibilidad técnica del sistema analítico.

Los resultados formales de estas pruebas se integran posteriormente en el capítulo de resultados mediante una tabla consolidada, con el fin de hacer explícita la contrastación de hipótesis y fortalecer la trazabilidad metodológica del estudio.

7.7.5. Ética y legalidad en la recolección y análisis

La recolección y el procesamiento de los datos se realizaron bajo principios de legalidad, transparencia y respeto por las condiciones de acceso a la información digital. El

corpus se construyó a partir de contenido publicado en fuentes abiertas y de acceso público, sin requerir autenticación, suscripción privada ni intervención sobre sistemas protegidos.

Asimismo, se siguieron principios éticos para investigación en internet orientados al uso responsable de información digital de acceso público y a la minimización de riesgos para los actores involucrados (Association of Internet Researchers, 2019).

La extracción automatizada se realizó respetando las directrices definidas en los archivos robots.txt de cada sitio web y, cuando estaban disponibles, las condiciones públicas de uso correspondientes. Se recolectó exclusivamente información accesible sin restricciones para agentes automatizados.

No se emplearon técnicas de evasión de controles técnicos, automatización agresiva ni manipulación de mecanismos de protección de los sitios.

En aquellos casos donde el acceso histórico a publicaciones del periodo 2022–2023 no se encontraba disponible en el sitio activo, se consultaron repositorios públicos de preservación digital, específicamente Internet Archive - Wayback Machine, accediendo a copias archivadas previamente disponibles para el público general. Este procedimiento permitió preservar la integridad temporal del corpus sin generar carga adicional sobre la infraestructura de los medios, garantizando además la reproducibilidad del corpus utilizado.

El estudio no involucró interacción con sujetos humanos, recopilación de datos personales identificables o sensibles, ni manipulación de información confidencial. En consecuencia, se clasifica como investigación sin riesgo, conforme a la normativa nacional vigente para investigación en salud, Resolución 8430 de 1993 del Ministerio de Salud y Protección Social de Colombia (Ministerio de Salud y Protección Social de Colombia, 1993).

El análisis automatizado se realizó a nivel agregado, sin identificación ni evaluación de individuos específicos, y con fines exclusivamente académicos. Los resultados se presentan en términos de patrones discursivos colectivos y no constituyen juicios sobre actores particulares.

Este enfoque reduce riesgos de interpretación indebida y se alinea con principios de responsabilidad en investigación computacional.

En conjunto, el diseño metodológico garantiza la coherencia del estudio con los principios éticos de investigación documental y el uso responsable de técnicas automatizadas aplicadas a información pública.

7.7.6. Métricas de recolección, control y calidad del corpus bruto

Con el fin de garantizar trazabilidad, reproducibilidad y control de calidad en la fase de captura de datos, se incorporó un conjunto explícito de métricas técnicas y descriptivas asociadas al corpus bruto, previas a las etapas de filtrado temático estricto y modelado discursivo. Esta documentación permite auditar el volumen inicial de información recolectada, las exclusiones técnicas aplicadas y la transición hacia el corpus analítico final.

Proceso de captura automatizada y cobertura temporal

La recolección se realizó mediante spiders especializados desarrollados en Scrapy, derivados de una clase base común (BaseSitemapSpider), lo que permitió unificar criterios de extracción y normalización de metadatos.

El proceso cubrió sistemáticamente el período enero de 2022 – diciembre de 2023. Cada documento incorporó:

- Fecha original de publicación (date).
- Marca temporal de captura (scraped_at).
- Identificación del spider y versión utilizada.

Esto permite auditoría temporal y reproducibilidad técnica.

Volumen bruto y preclasificación temática

Durante la fase inicial se procesaron 6.133 artículos, de los cuales 704 superaron la validación preliminar (filtrado suave basado en detección léxica de términos asociados a salud y corrupción).

Este conjunto preclasificado no constituye aún el corpus analítico final. Posteriormente, mediante:

- Intersección temática estricta SALUD ∩ CORRUPCIÓN.
- Eliminación de duplicados (Bloom Filter).
- Validación temática definitiva.

Se conformó el corpus analítico final de 518 documentos.

Esta diferenciación explícita permite separar exclusiones técnicas de decisiones metodológicas posteriores.

Métricas estructurales del corpus bruto

Para cada artículo se registraron variables estructurales que permiten caracterizar el corpus previo a depuración:

- Medio de origen.
- Fecha de publicación.
- Categoría editorial.
- Número de palabras (word_count).
- Número de caracteres (char_count).
- Clasificación de longitud textual.

Estas métricas permiten identificar posibles sesgos de concentración por medio, formato o periodo temporal, asegurando heterogeneidad estructural del corpus bruto.

Métricas técnicas del proceso de scraping

El sistema registró métricas operacionales asociadas al desempeño de extracción, incluyendo:

- Número total de solicitudes HTTP.
- Distribución de códigos de respuesta (200, 3xx, 4xx, 5xx).
- Volumen total de datos transferidos.
- Tiempo total de ejecución.
- Eventos de reintento.

Estas métricas permiten evaluar estabilidad técnica, eficiencia operativa y posibles limitaciones estructurales de cada fuente.

Asimismo, estos registros permiten auditar el cumplimiento de buenas prácticas de acceso automatizado y detectar posibles restricciones operativas impuestas por los servidores.

Relación entre corpus bruto y corpus analítico final

La documentación del volumen inicial (6.133 artículos), la preclasificación técnica (704 registros) y el corpus final validado (518 documentos) evidencia que la muestra utilizada en el análisis discursivo es resultado de un proceso controlado, reproducible y metodológicamente transparente.

Tabla 15. Valores brutos de recolección y resultados de filtrado por medio

Medio	Artículos procesados	Artículos aceptados	% aceptación	Solicitudes HTTP	Respuestas 200	Respuestas 404	Datos transferidos (MB)	Tiempo total (min)
Revista Cambio	1.136	120	10,6 %	1.414	1.197	4	39,1	50
Cuestión Pública	1.129	130	11,5 %	12	12	0	11,4	1,7
Las 2 Orillas	181	15	8,3 %	226	225	1	14,3	4,7
La Silla Vacía	232	46	19,8 %	312	264	24	27,6	18,5
Pulzo	3.294	380	11,5 %	3.469	3.333	0	311,8	106,2
Revista Metro	161	13	8,1 %	192	192	0	4,8	5,9
Total	**6.133**	**704**	**11,5 %***	**5.625**	**5.223**	**29**	**408,9**	**187,0**

Nota. El porcentaje de aceptación corresponde a la razón entre artículos aceptados y artículos procesados por medio. El total de 704 artículos aceptados corresponde a la fase de preclasificación técnica; el corpus analítico final se compone de 518 documentos tras el filtrado temático estricto y la deduplicación.

*El porcentaje total corresponde al cociente entre el total de artículos aceptados (704) y el total procesado (6.133).

Nota adicional. El número de solicitudes HTTP puede ser inferior al número de artículos procesados debido a los accesos mediante sitemap, recuperación estructurada de contenidos (por ejemplo, endpoints o archivos indexados) y reutilización de respuestas previamente obtenidas durante la exploración del sitio.

En conjunto, el diseño metodológico descrito desde la arquitectura de recolección automatizada hasta la validación estadística del modelado temático y el análisis de sentimiento establece un marco analítico reproducible, transparente y coherente con los objetivos e hipótesis planteados. Con base en este esquema metodológico y una vez consolidado el corpus analítico final de 518 documentos, el capítulo siguiente presenta los resultados empíricos derivados del modelado temático, el análisis de tono emocional y los cruces comparativos entre medios digitales, orientados a contrastar las hipótesis de investigación.

8. Trabajo de Campo

El desarrollo empírico de la investigación se estructuró conforme al enfoque CRISP-DM, operacionalizado como un pipeline computacional reproducible que integra la selección de fuentes, la recolección automatizada de datos, la depuración textual, el modelado temático, el análisis de sentimiento, la identificación de actores y la validación estadística de resultados. Esta sección describe la implementación técnica del proceso analítico y las decisiones metodológicas que condicionan la construcción del corpus y el alcance interpretativo del estudio.

8.1. Fases metodológicas

A continuación, se presenta el esquema de trabajo adoptado, sintetizado en la Tabla 16, que resume las fases metodológicas, los objetivos operativos y las tecnologías empleadas en el proyecto.

Tabla 16. Fases metodológicas, objetivos y herramientas tecnológicas aplicadas al proyecto

Fase	Objetivo	Tecnologías y herramientas empleadas
1. Selección de medios	Definir una muestra analítica de medios digitales (nacionales y regionales).	Criterios editoriales, métricas de presencia digital (SEO) y verificación técnica de accesibilidad web.
		Métricas de autoridad de dominio (SEO).
		Verificación manual de accesibilidad web.
2. Recolección del corpus	Extraer masivamente el archivo histórico de noticias (2022-2023).	Python + Scrapy: Framework principal de extracción (Spiders).
		Requests/BeautifulSoup: Para manejo de excepciones puntuales.
		Salida: Archivos JSON estructurados por medio.
3. Ingeniería de Datos	Unificar, limpiar y normalizar el texto para el modelado.	Pandas: Manipulación de dataframes y formato Parquet (alta eficiencia).
		SpaCy/NLTK: Tokenización, eliminación de stopwords y lematización.
		Expresiones regulares (regex): Limpieza de ruido HTML y caracteres especiales.
4. Análisis computacional (PLN)	Aplicar modelos de PLN para identificar estructuras temáticas, polaridad emocional y actores	Modelado de tópicos: Gensim (LDA Multicore)
		Modelo contrastivo: BERTopic (comparación estructural del espacio temático)

	discursivos en el corpus periodístico.	Modelo no paramétrico: HDP (Hierarchical Dirichlet Process)
		Análisis de sentimiento: pysentimiento (RoBERTa)
		Reconocimiento de entidades: spaCy (NER)
		Orquestación: main_pipeline.py
5. Visualización de Datos	Traducir los resultados numéricos en representaciones gráficas interpretables.	Matplotlib / Seaborn: Generación de mapas de calor, barras y líneas de tiempo (PNG estáticos).
		PyLDAvis: Visualización interactiva de distancia inter-tópicos.
		WordCloud: Nubes de palabras por tópico.
6. Validación del modelo	Verificar la robustez matemática y la coherencia semántica de los hallazgos obtenidos mediante modelado temático y análisis complementarios.	Grid Search: optimización de hiperparámetros del modelo LDA.
		Métricas de coherencia: C_v y C_npmi (selección del número óptimo de tópicos).
		Diversidad temática (Topic Diversity): evaluación de redundancia semántica.
		Validación comparativa: contraste estructural con modelos alternativos (HDP y BERTopic).
		Métricas de concordancia intermodelo: NMI, ARI y Jaccard top-words.
		Análisis de estabilidad: sensibilidad a parámetros, consistencia temporal y revisión de documentos representativos por tópico.
7. Sistematización	Garantizar la reproducibilidad del código y la estructura del proyecto.	Entorno Virtual: Gestión de dependencias (requirements.txt).
		Estructura Modular: Separación de scripts (src/) y datos (data/).
		Git/GitHub: Control de versiones del código fuente.
8. Interpretación Final	Integrar los hallazgos computacionales con la teoría de análisis de discurso.	Redacción técnica del informe de tesis.
		Interpretación discursiva de documentos representativos por tópico.

Nota. La tabla sintetiza las fases del proceso metodológico, los objetivos operativos y las tecnologías empleadas en el desarrollo del proyecto, conforme a la operacionalización del enfoque CRISP-DM aplicada al análisis de discurso mediante Procesamiento de Lenguaje Natural (PLN).

8.2. Selección y depuración de fuentes de información

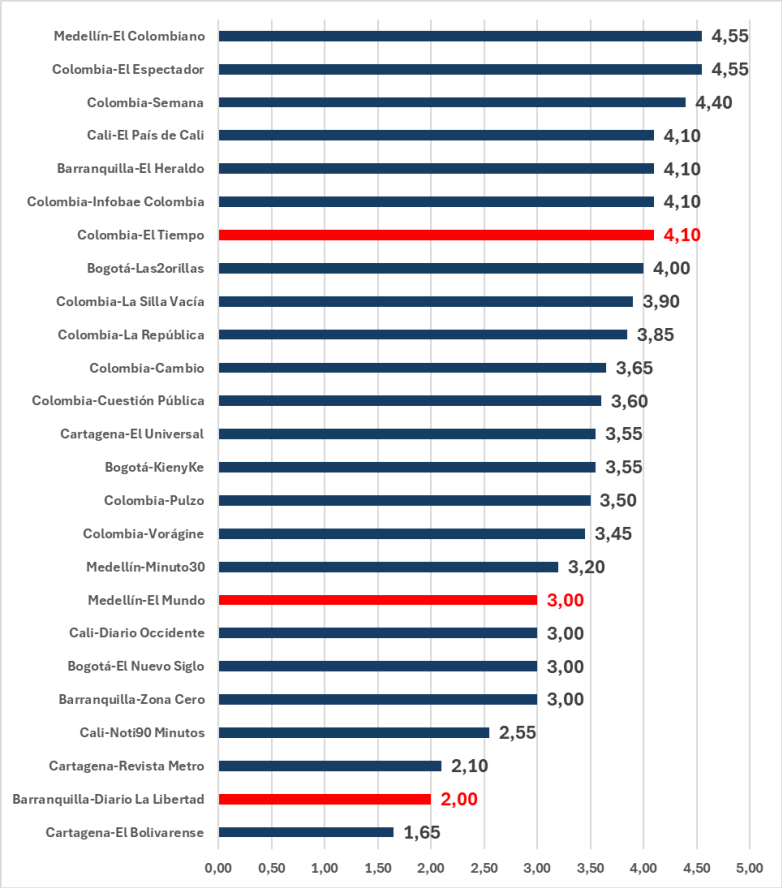
La conformación del corpus se ejecutó en dos etapas secuenciales:

(i) una preselección basada en métricas de presencia digital y relevancia investigativa;

(ii) una auditoría de viabilidad técnica para la extracción retrospectiva del periodo 2022–2023, evaluada conforme a criterios de accesibilidad automatizada y cumplimiento ético-legal.

En esta fase no se realiza aún la selección definitiva de medios, sino la identificación de aquellos portales cuya arquitectura y políticas de acceso permiten la recolección automatizada sistemática y reproducible. Los medios que superan esta verificación permanecen como candidatos elegibles para la etapa posterior de selección de la muestra definitiva.

Figura 6. Evaluación de viabilidad técnica para extracción automatizada según políticas de acceso



Nota. Elaborada a partir de los puntajes de evaluación (Tabla 14). Los medios en azul representan portales técnicamente viables para la extracción automatizada conforme a sus políticas de acceso público y archivos robots.txt, mientras que los medios en rojo corresponden a sitios con restricciones explícitas de acceso que impiden la recolección automatizada bajo los criterios éticos y técnicos del estudio.

Este procedimiento garantiza que la exclusión de fuentes responda exclusivamente a limitaciones técnicas verificables y no a consideraciones editoriales o temáticas.

8.2.1. Conformación del Corpus Operativo

Aunque el diseño metodológico definió un marco analítico inicial de diez medios digitales como universo de referencia, la recolección automatizada retrospectiva del periodo 2022–2023 se operacionalizó finalmente sobre seis medios que cumplieron simultáneamente los criterios de accesibilidad técnica, disponibilidad histórica y viabilidad de extracción automatizada. Esta submuestra constituye el corpus operativo efectivo del estudio y responde a restricciones técnicas verificables durante la fase de implementación, sin comprometer los objetivos analíticos ni la estrategia de comparación entre categorías de medios.

La selección del corpus final se ejecutó mediante un pipeline secuencial que integró filtrado temático estricto SALUD ∩ CORRUPCIÓN, limpieza estructural, eliminación de boilerplate y deduplicación. Como resultado, el corpus analizado quedó conformado por 518 artículos periodísticos (2022–2023).

Tabla 17. Caracterización técnica del corpus final por tipo de medio (nacional y regional)

Medio	Alcance / Enfoque	Configuración Técnica y Estrategia de Extracción
Cuestión Pública	**Nacional** (Investigación / Datos)	**Estrategia:** Acceso programático a endpoints públicos accesibles mediante inspección técnica. **Fuente:** Endpoints JSON directos (/wp-json/wp/v2/posts). **Ventaja:** Obtención de datos crudos sin ruido visual.
Revista Cambio	**Nacional** (Política / Poder)	**Estrategia:** Extracción de datos estructurados generados por aplicaciones Next.js **Fuente:** Extracción de objetos JSON del Build Manifest. **Acción:** Limpieza de parámetros de rastreo (utm_) para cumplimiento estricto.
La Silla Vacía	**Nacional / Regional** (Análisis de Poder)	**Estrategia:** Navegación por Archivos Mensuales. **Fuente:** Directorios de fecha (/2022/01/) para evitar bloqueos del buscador. **Configuración:** Configuración de retardo entre solicitudes (3 segundos) para mitigar errores 429.
Pulzo	**Nacional** (Agregador / Volumen)	**Estrategia:** Selectores jerárquicos CSS/XPath. **Fuente:** Sitemaps XML sectorizados (/nacion, /salud). **Acción:** Lógica de *fallback* para capturar contenido en estructuras HTML variables.
Las 2 Orillas	**Bogotá / Ciudadanía** (Opinión)	**Estrategia:** Reconstrucción de URLs de Archivo. **Fuente:** Generación algorítmica de rutas históricas. **Enfoque:** Recolección masiva de denuncias ciudadanas.
Revista Metro	**Cartagena / Caribe** (Local)	**Estrategia:** Navegación por archivos históricos de WordPress y normalización lingüística. **Fuente:** Navegación directa por carpetas mensuales. **Acción:** Procesamiento de Lenguaje Natural para normalizar fechas textuales.

Nota. La tabla presenta los medios que superaron la auditoría técnica de accesibilidad y cuyas arquitecturas web permiten la recuperación automatizada del contenido histórico correspondiente al periodo 2022–2023.

La heterogeneidad técnica de las fuentes implicó exigencias de limpieza diferenciadas durante la fase de ingeniería de datos. Mientras Cuestión Pública permitió la recuperación de datos estructurados mediante endpoints JSON directos, con mínima interferencia de ruido visual, Pulzo requirió extracción sobre HTML variable mediante selectores jerárquicos CSS/XPath y lógica de fallback, debido a la presencia de componentes editoriales y publicitarios integrados en la estructura de página. Esta diferencia se reflejó también en la carga operativa del proceso de scraping, como se observa en las métricas técnicas reportadas en la Tabla 15.

Criterios de delimitación: Adicionalmente, medios preseleccionados como Vorágine (Nacional), Diario Occidente (Cali), El Nuevo Siglo (Bogotá) y El Bolivarense (Cartagena) no fueron incluidos en la fase de extracción masiva. Esta decisión se soportó en un criterio cualitativo de delimitación operativa basado en la redundancia temática esperada frente a los ejes discursivos ya cubiertos por las fuentes incluidas en el corpus operativo. En particular, los medios finalmente seleccionados ya representaban combinaciones diferenciadas de alcance nacional y regional, periodismo investigativo, análisis político, agregación de alto volumen y cobertura local, por lo que la incorporación de nuevas fuentes con perfiles parcialmente solapados ofrecía un valor analítico incremental limitado frente al costo computacional y técnico de su extracción. Su exclusión permitió concentrar los recursos en portales con mayor disponibilidad histórica y estabilidad técnica, sin afectar de manera sustantiva la diversidad analítica del corpus final.

8.2.2. Exclusiones técnicas y limitaciones de acceso

Durante la fase de validación de los spiders (rastreadores), se identificaron barreras tecnológicas críticas en la arquitectura web de siete (7) portales informativos inicialmente preseleccionados, lo que imposibilitó la recolección automatizada de datos bajo los parámetros de ética, legalidad, viabilidad técnica y reproducibilidad definidos en el proyecto. En consecuencia, el corpus operativo se restringió a medios con acceso abierto o semiautomatizable, privilegiando la trazabilidad del proceso sobre una cobertura potencialmente más amplia, pero metodológicamente menos robusta.

A continuación, se detalla la justificación técnica para la exclusión de cada medio:

Tabla 18. Medios excluidos por inviabilidad técnica

Medio	Ámbito de cobertura	Barrera Principal Detectada
El Espectador	Nacional	Renderizado dinámico (CSR) / Protección ARC Publishing.
Revista Semana	Nacional	*Paywall* rígido / Sitemaps rodantes (Rolling sitemaps).
Infobae	Nacional	Dependencia total de JavaScript para listados.
Minuto30	Medellín	Sesgo de recencia / Archivos inaccesibles.
El Colombiano	Antioquia	Sitemaps bloqueados / Inaccesibilidad histórica.
El País	Cali	Errores de Endpoints (404) / Renderizado dinámico.
El Heraldo	Barranquilla	Renderizado dinámico (CSR) / Ofuscación de metadatos.

Nota. La tabla resume los hallazgos de la revisión técnica realizada a los portales informativos preseleccionados, detallando las barreras de arquitectura web (renderizado dinámico, ofuscación de datos o restricciones de acceso) que impidieron la indexación automatizada del contenido histórico para el periodo 2022-2023.

Justificación técnica:

Las barreras identificadas responden a características específicas de la arquitectura web de cada medio, que impiden la recuperación automatizada retrospectiva bajo condiciones éticas y técnicas establecidas en el diseño del estudio. En particular:

- **Renderizado dinámico del lado del cliente (CSR)** en plataformas como ARC Publishing (El Espectador y El Heraldo), donde el contenido textual se genera mediante JavaScript en el navegador y no está disponible en las respuestas HTTP iniciales.
- **Sistemas de acceso restringido o paywall combinados con sitemaps dinámicos** (Revista Semana), que impiden obtener un archivo histórico completo y estable del periodo analizado.
- **Dependencia total de JavaScript para la generación de listados de noticias** (Infobae), lo que requeriría el uso de navegadores headless u otras técnicas avanzadas fuera del alcance operativo del proyecto.
- **Ausencia o inaccesibilidad de archivos históricos indexables** (El Colombiano y Minuto30), lo que imposibilita la recolección sistemática retrospectiva.
- **Errores persistentes en endpoints y renderizado dinámico sin segmentación temporal accesible** (El País), que impiden aislar el contenido correspondiente al periodo 2022–2023.

En conjunto, estas limitaciones técnicas impidieron la construcción de procedimientos de recolección automatizada reproducibles y conformes con los estándares éticos y técnicos del estudio, motivo por el cual dichos medios fueron excluidos del corpus operativo.

8.2.3. Estrategia de mitigación y representatividad

Con el fin de mitigar la reducción de cobertura territorial derivada de la exclusión técnica de varios diarios regionales tradicionales, se incorporaron medios digitales con capacidad de cobertura nacional descentralizada o enfoque regional explícito. Esta estrategia busca preservar la diversidad geográfica del corpus y reducir posibles sesgos derivados de la disponibilidad técnica de las fuentes.

- **La Silla Vacía:** Se incorpora como fuente estratégica para el análisis regional gracias a sus capítulos especializados ("Silla Caribe", "Silla Pacífico", "Silla Paisa"), permitiendo recuperar la perspectiva regional sobre asuntos públicos y políticas sectoriales.
- **Las 2 Orillas:** Su modelo de "orillas" permite capturar el contenido de participación y opinión ciudadana desde las regiones, mitigando parcialmente la pérdida de cobertura de medios regionales tradicionales.
- **Revista Metro:** Se mantiene como fuente primaria para la región Caribe (Cartagena y Bolívar), garantizando un punto de vista local específico sobre la corrupción en salud en una de las zonas más afectadas.

Esta reconfiguración del corpus permite mantener una perspectiva analítica plural y territorialmente diversa sobre el fenómeno de la corrupción en salud, a pesar de las restricciones técnicas de acceso a algunos portales informativos tradicionales.

El corpus final no es arbitrario, sino condicionado por restricciones técnicas verificables.

8.3. Análisis descriptivo del corpus

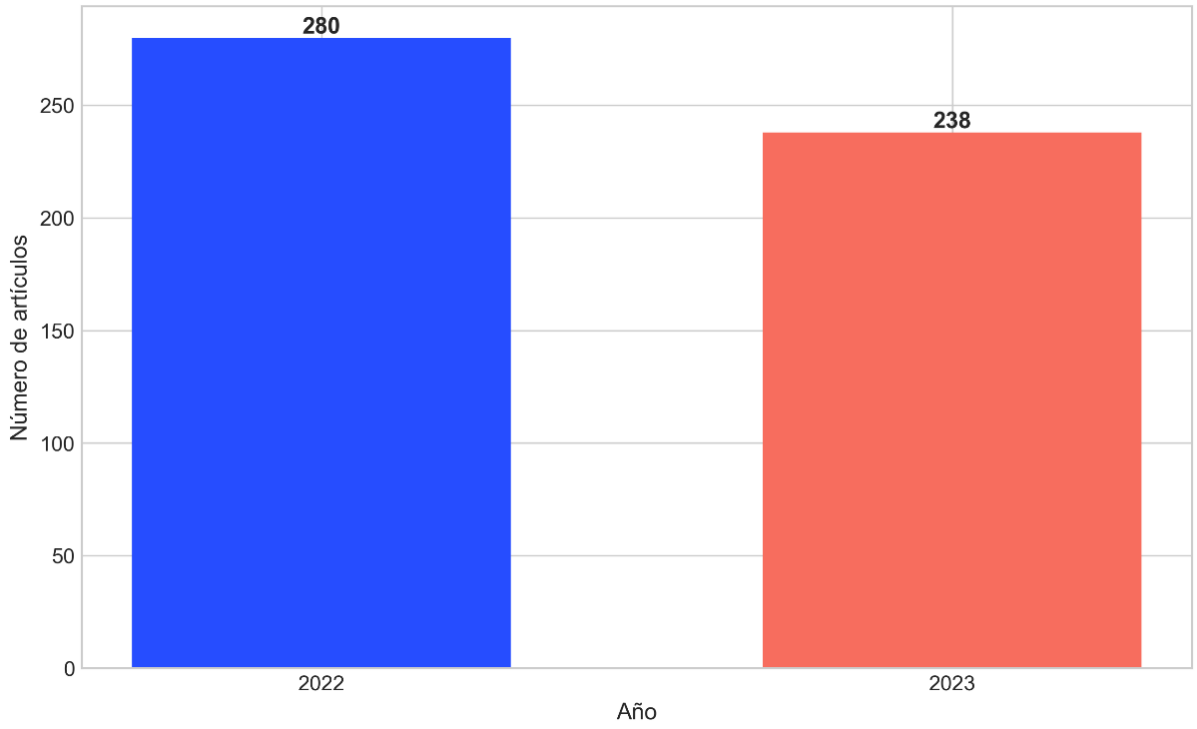
El análisis descriptivo del corpus tiene como propósito caracterizar empíricamente la muestra antes de la aplicación de los modelos de Procesamiento de Lenguaje Natural (PLN), evaluando su distribución temporal, editorial, geográfica y estructural. Esta fase permite verificar la suficiencia analítica del conjunto de datos y contextualizar los resultados obtenidos en las etapas posteriores.

Tras el filtrado temático estricto, la limpieza estructural, la eliminación de contenido boilerplate y la deduplicación, el corpus final quedó conformado por 518 artículos periodísticos publicados entre enero de 2022 y diciembre de 2023, constituyendo la base empírica del estudio.

8.3.1. Distribución temporal y evolución del volumen informativo

El análisis temporal permite identificar la intensidad de la cobertura mediática sobre la corrupción en el sector salud durante el periodo analizado. La Figura 7 muestra la distribución anual de los artículos recuperados.

Figura 7. Distribución anual de artículos recuperados (2022–2023)

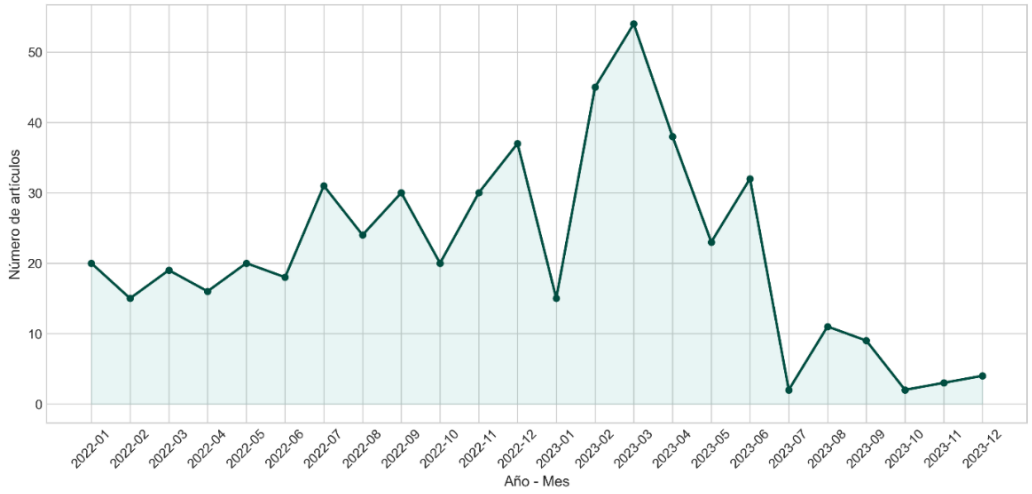


Nota. Elaborada a partir del corpus analítico.

El corpus registra un mayor volumen de publicaciones en 2022 (280 artículos) frente a 2023 (238 artículos), lo que representa una reducción de 42 artículos (≈15 %). Esta diferencia sugiere que la visibilidad mediática del fenómeno se mantuvo sostenida en ambos años, aunque con menor intensidad relativa en el segundo periodo.

Para examinar con mayor detalle la dinámica temporal, la Figura 8 presenta la evolución mensual del volumen informativo.

Figura 8. Evolución mensual del volumen informativo

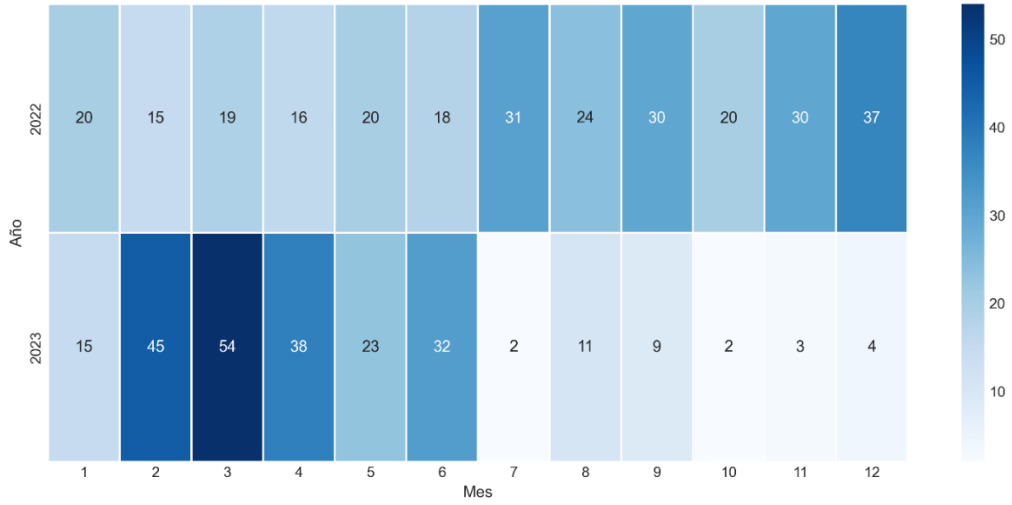


Nota. Elaborada a partir del corpus analítico.

Durante 2022, la cobertura muestra un comportamiento relativamente estable, con fluctuaciones moderadas a lo largo del año y valores mensuales comprendidos principalmente entre 15 y 37 artículos. En contraste, 2023 presenta una dinámica más volátil, caracterizada por un aumento significativo en los primeros meses especialmente entre febrero y marzo seguido de una disminución progresiva y pronunciada durante el segundo semestre.

Con el fin de sintetizar comparativamente esta variación, la Figura 9 presenta un mapa de calor de la intensidad de publicación mensual.

Figura 9. Intensidad de publicación mensual (mapa de calor)



Nota. Elaborada a partir del corpus analítico.

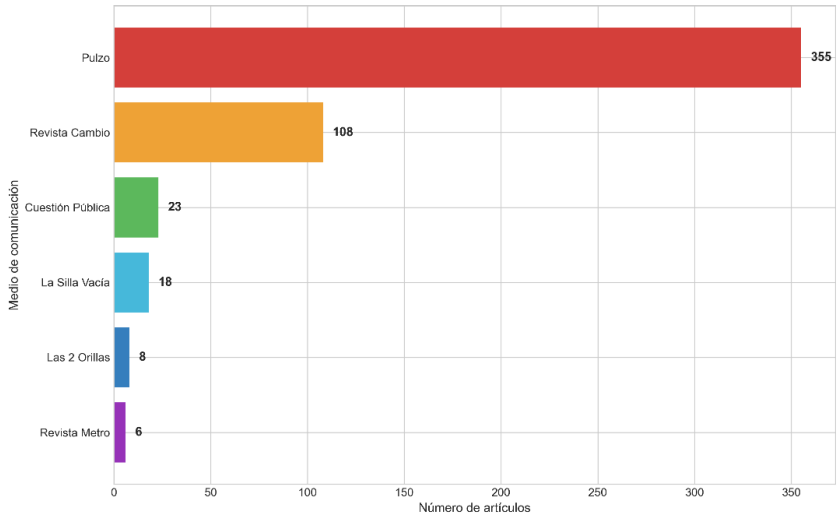
El mapa de calor confirma la diferencia estructural entre ambos años. Mientras que 2022 exhibe una distribución relativamente homogénea de la cobertura, 2023 concentra la mayor intensidad informativa en el primer trimestre, particularmente entre febrero y abril, seguido de una caída sostenida del volumen de publicaciones a partir de julio.

Este patrón indica que la atención mediática sobre la corrupción en el sector salud no se distribuye de manera uniforme en el tiempo, lo que sugiere un comportamiento no uniforme en la cobertura mediática, caracterizado por picos temporales de alta intensidad informativa.

8.3.2. Distribución editorial y dinámicas de publicación

El análisis por medio de comunicación permite evaluar la contribución relativa de cada plataforma informativa al corpus y detectar posibles concentraciones editoriales que influyan en la cobertura del fenómeno estudiado. La Figura 10 presenta la distribución total de artículos por medio durante el periodo 2022–2023.

Figura 10. Participación cuantitativa por medio de comunicación



Nota. Elaborada a partir del corpus analítico.

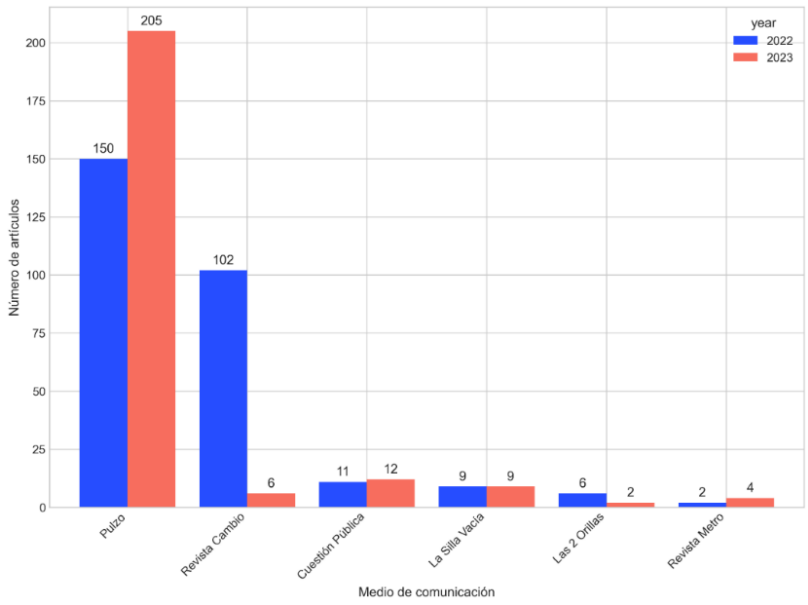
Los resultados evidencian una fuerte concentración del corpus en Pulzo (355 artículos), que representa cerca del 70 % del total analizado. En segundo lugar, se ubica Revista Cambio (108 artículos), mientras que los demás medios presentan contribuciones considerablemente menores: Cuestión Pública (23), La Silla Vacía (18), Las 2 Orillas (8) y Revista Metro (6).

Esta distribución refleja diferencias sustantivas en los ritmos de publicación, las estrategias editoriales, la frecuencia editorial y el volumen de producción noticiosa dentro del ecosistema digital. Asimismo, indica que la cobertura del tema no se distribuye de manera uniforme entre medios, lo que introduce un potencial sesgo de representatividad asociado al volumen de publicación.

Aunque Pulzo concentra la mayor proporción de documentos del corpus, los resultados no se interpretan a partir de frecuencias absolutas aisladas, sino mediante distribuciones relativas y contrastes entre medios. En este sentido, la variación temática observada por medio indica que los tópicos identificados no reproducen de forma mecánica la línea editorial de un único actor, aunque sí exigen cautela en la interpretación de la representatividad global del ecosistema mediático analizado.

La Figura 11 compara la contribución de cada medio entre los años 2022 y 2023.

Figura 11. Dinámica comparativa de publicación por medio (2022–2023)



Nota. Elaborada a partir del corpus analítico.

El contraste interanual muestra comportamientos claramente diferenciados. Pulzo incrementa de forma significativa su volumen de publicaciones en 2023 (de 150 a 205 artículos), consolidándose como el medio dominante en términos de volumen informativo. En

sentido opuesto, Revista Cambio presenta una reducción marcada, pasando de 102 artículos en 2022 a solo 6 en 2023.

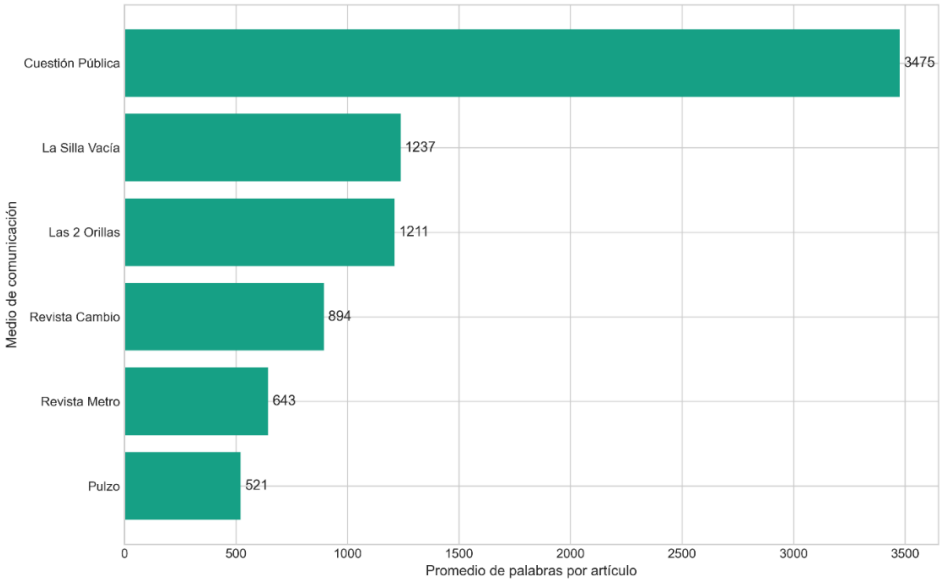
Los medios Cuestión Pública y La Silla Vacía mantienen niveles relativamente estables entre ambos años, mientras que Las 2 Orillas registra una disminución moderada y Revista Metro una participación marginal con ligeras variaciones.

En conjunto, estos resultados evidencian una heterogeneidad estructural en la producción de contenidos, caracterizada por una alta concentración en un número reducido de medios y por dinámicas interanuales disímiles. Dado este desequilibrio, los análisis posteriores se basan en medidas relativas y procedimientos de normalización que permiten examinar las características discursivas sin que el volumen absoluto de publicaciones determine de forma directa la interpretación de los resultados.

8.3.3. Caracterización del contenido: Longitud y Profundidad

La longitud de los artículos constituye un indicador estructural relevante para evaluar la densidad informativa y el grado de desarrollo narrativo del corpus. La Figura 12 presenta el promedio de palabras por artículo para cada medio analizado.

Figura 12. Promedio de palabras por artículo por medio de comunicación



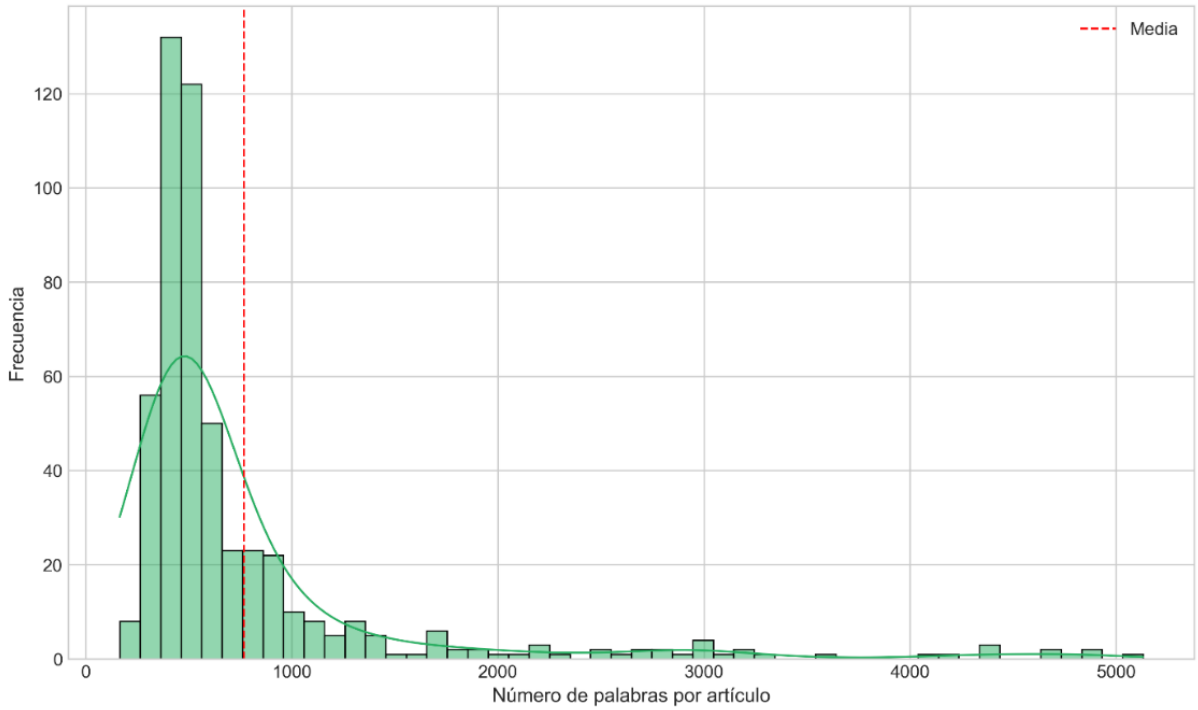
Nota. Elaborada a partir del corpus analítico.

Los resultados evidencian diferencias sustantivas en la extensión promedio de los textos entre medios. Cuestión Pública presenta la mayor longitud media, con aproximadamente 3475 palabras por artículo, muy por encima del resto del corpus. Le siguen La Silla Vacía y Las 2 Orillas, con promedios cercanos a 1200 palabras, mientras que Revista Cambio presenta una extensión intermedia (alrededor de 900 palabras).

En contraste, Revista Metro y Pulzo exhiben los textos más breves, con promedios inferiores a 700 y 550 palabras respectivamente. Estas variaciones reflejan diferencias en los formatos editoriales y en el tipo de contenido publicado, desde reportajes extensos hasta piezas informativas más concisas.

A nivel global, la Figura 13 muestra la distribución de la longitud de los artículos mediante un histograma que incorpora la media del corpus como referencia.

Figura 13. Distribución de longitud de los artículos del corpus



Nota. Elaborada a partir del corpus analítico.

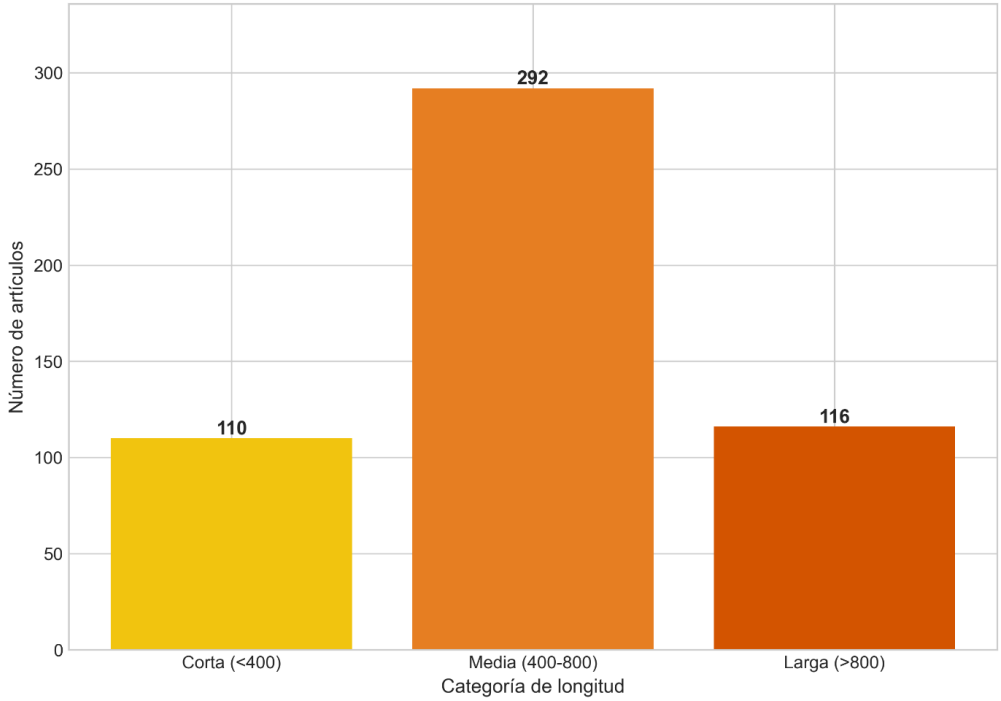
La distribución presenta una marcada asimetría positiva, con una alta concentración de textos en rangos de longitud baja y media y una cola extendida asociada a un número reducido

de artículos considerablemente más extensos. La mayor parte de las publicaciones se sitúa aproximadamente entre 300 y 900 palabras, mientras que algunos documentos superan ampliamente este rango, alcanzando varios miles de palabras.

Esta estructura indica la presencia de valores extremos que incrementan la media del corpus sin alterar el predominio de textos de extensión moderada.

Para sintetizar esta estructura, la Figura 14 clasifica los artículos en tres categorías de longitud: corta (<400 palabras), media (400–800 palabras) y larga (>800 palabras).

Figura 14. Clasificación de artículos por categoría de longitud



Nota. Elaborada a partir del corpus analítico.

Los resultados muestran un predominio de textos de longitud media (292 artículos), que representan más de la mitad del corpus, seguidos por artículos largos (116) y cortos (110), con proporciones similares entre estos dos últimos grupos.

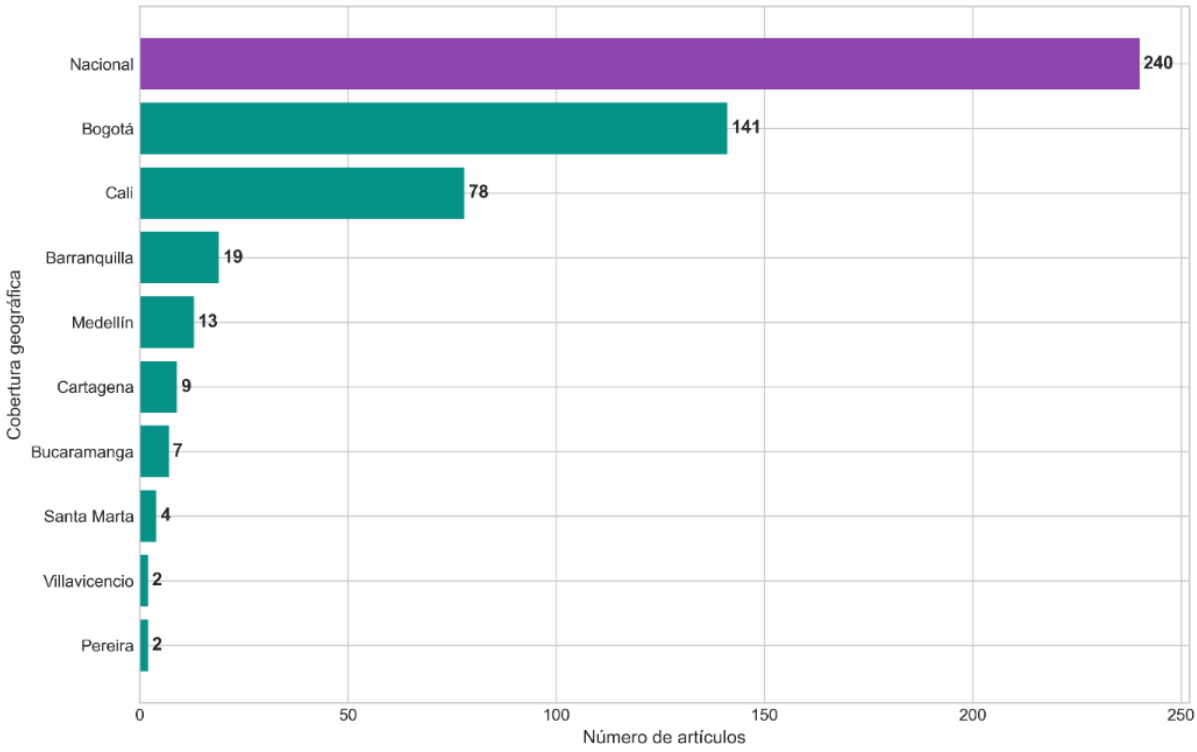
Esta distribución indica que el corpus combina piezas informativas breves con contenidos de mayor extensión analítica, sin que ninguno de los extremos domine de manera absoluta.

En conjunto, el corpus presenta una variabilidad significativa en la extensión de los textos, lo que favorece la diversidad informativa y proporciona suficiente material lingüístico para la identificación de patrones temáticos y discursivos mediante técnicas de Procesamiento de Lenguaje Natural.

8.3.4. Distribución geográfica del enfoque informativo

El análisis del ámbito geográfico permite evaluar la distribución territorial de la cobertura mediática sobre corrupción en el sector salud. La Figura 15 presenta la frecuencia de artículos según el nivel geográfico principal al que hace referencia cada publicación.

Figura 15. Distribución geográfica del enfoque de la noticia



Nota. Elaborada a partir del corpus analítico.

Los resultados evidencian un claro predominio del enfoque nacional (240 artículos), seguido por Bogotá (141). Entre las coberturas regionales destaca Cali (78), por otra parte, ciudades como Barranquilla (19) y Medellín (13) presentan participaciones significativamente menores, mientras que el resto de los territorios registra contribuciones marginales.

Esta distribución sugiere una marcada concentración de la cobertura en el ámbito nacional y en la capital, aunque la presencia de casos regionales permite incorporar referencias territoriales específicas. En conjunto, el corpus combina narrativas de alcance nacional con coberturas locales puntuales, lo que posibilita examinar tanto dimensiones estructurales del fenómeno como manifestaciones territoriales particulares.

8.3.5. Consideraciones finales sobre la muestra

El análisis descriptivo permite establecer la estructura general del corpus, evidenciando variaciones en volumen de publicación, extensión de los textos, cobertura geográfica y participación editorial entre los medios analizados. Estas diferencias reflejan las dinámicas propias del ecosistema mediático digital colombiano y no constituyen artefactos derivados del proceso de selección, sino características empíricas del conjunto de fuentes considerado.

El corpus combina contenidos de alta frecuencia informativa con artículos de mayor desarrollo analítico, lo que introduce diversidad en la estructura discursiva del material y permite examinar distintas modalidades de tratamiento del fenómeno de la corrupción en salud.

Desde el punto de vista metodológico, esta configuración resulta adecuada para la aplicación de técnicas de Procesamiento de Lenguaje Natural orientadas a identificar patrones semánticos, temáticos y emocionales, al proporcionar suficiente variabilidad textual y volumen de datos para el modelado automatizado. En consecuencia, el corpus constituye una base empírica consistente para las fases posteriores del análisis computacional del discurso.

8.4. Configuración y validación de los modelos temáticos

La validación del modelado temático se realizó mediante un enfoque multcriterio basado en métricas de coherencia semántica e indicadores de diferenciación temática. En particular, se emplearon la coherencia C_v, ampliamente utilizada por su correlación con la interpretabilidad humana de los tópicos, y la coherencia C_npmi, basada en información mutua normalizada que evalúa la consistencia de co-ocurrencia léxica entre términos.

Adicionalmente, se incorporó la métrica de diversidad temática (Topic Diversity) para estimar el grado de diferenciación entre los conjuntos de términos dominantes de cada tópico.

Este esquema permitió seleccionar configuraciones que equilibran interpretabilidad, coherencia semántica y separación estructural, evitando depender de un único indicador cuantitativo.

8.4.1. Diseño experimental y parámetros de modelado

El diseño experimental del modelado temático partió del corpus definitivo de 518 artículos previamente depurado y normalizado. Sobre esta base se construyó una representación léxica de bolsa de palabras (Bag-of-Words, BoW) asociada a un diccionario derivado del corpus, compatible con modelos probabilísticos no supervisados como LDA y HDP, incorporando el preprocesamiento lingüístico realizado en las fases anteriores del pipeline.

Se implementaron tres enfoques complementarios con funciones diferenciadas dentro del proceso analítico. El modelo LDA se adoptó como núcleo paramétrico principal, orientado a identificar una estructura temática interpretable mediante la especificación explícita del número de tópicos. Con este fin, se evaluó un rango de configuraciones con distintos valores de k, seleccionando la solución óptima en función de la coherencia semántica, la diversidad temática, la estabilidad del modelo y su comportamiento de generalización, evitando una fragmentación excesiva del espacio discursivo.

El modelo HDP se empleó como referencia no paramétrica para explorar la tendencia natural del corpus respecto al número de temas, sin imponer un valor predeterminado. Su función fue orientativa dentro del proceso de selección de k, no competitiva en términos de desempeño.

Por su parte, BERTopic se utilizó como contraste basado en representaciones semánticas densas derivadas de modelos de lenguaje preentrenados, con el propósito de evaluar la granularidad temática y la correspondencia estructural respecto a la solución obtenida mediante LDA.

Este diseño permitió integrar evidencia procedente de enfoques metodológicos complementarios, manteniendo a LDA como modelo principal y utilizando HDP y BERTopic como instrumentos de validación exploratoria y triangulación metodológica.

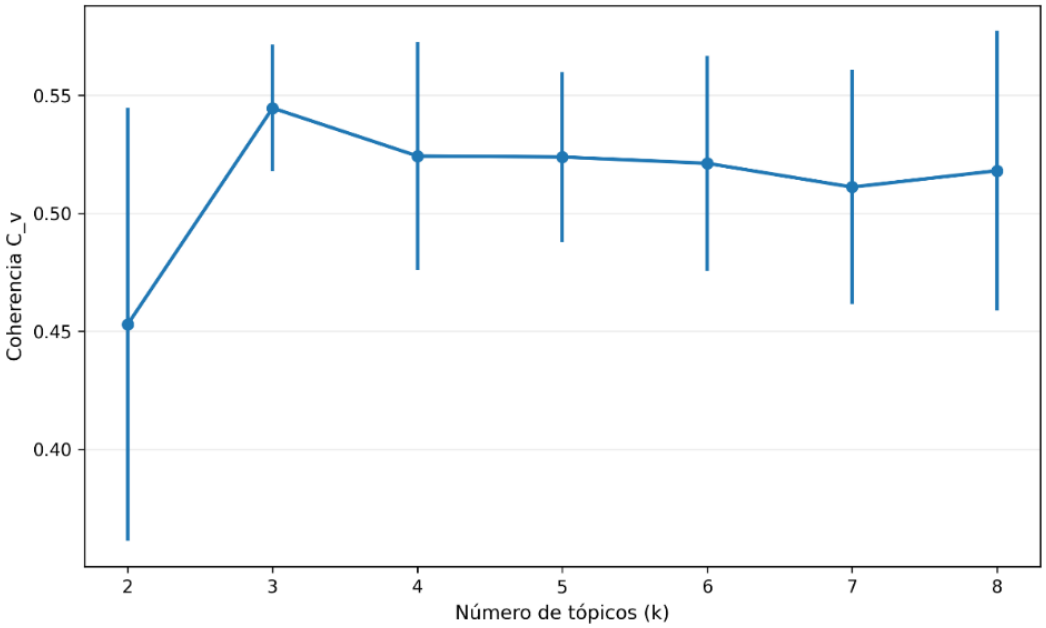
8.4.2. Optimización del modelo LDA

La optimización del modelo LDA se llevó a cabo mediante la evaluación sistemática de distintos valores del número de tópicos (k), considerando simultáneamente métricas de coherencia semántica, diversidad temática y desempeño en validación medido mediante log-perplejidad.

Los resultados muestran que la configuración con k = 3 ofrece el mejor equilibrio entre interpretabilidad, parsimonia y consistencia analítica. Esta solución alcanzó un valor de coherencia C_v = 0.5379, acompañado de niveles altos de diversidad temática y de estabilidad estructural, lo que indica una adecuada separación entre los ejes discursivos identificados sin incurrir en fragmentación excesiva.

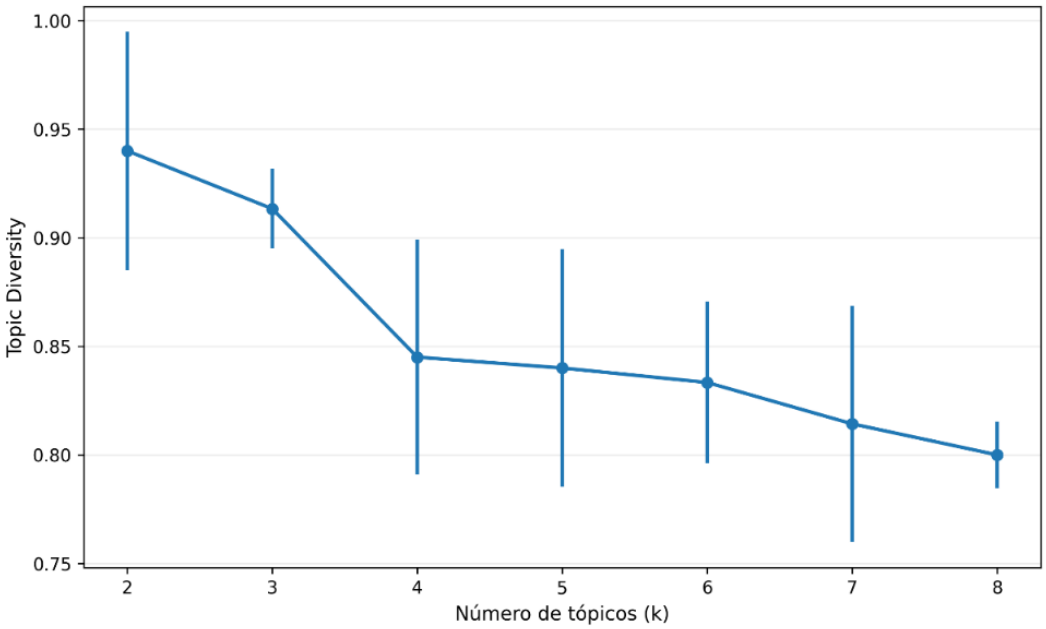
El análisis por valores de k evidencia que, aunque configuraciones con más tópicos mantienen niveles aceptables de coherencia, lo hacen a costa de una reducción progresiva de la diversidad temática y de una mayor especialización de los tópicos. En este contexto, la solución de tres tópicos resulta más adecuada para la interpretación sustantiva del corpus, al condensar el discurso en ejes reconocibles y diferenciados.

Figura 16. Coherencia temática C_v del modelo LDA en la validación final



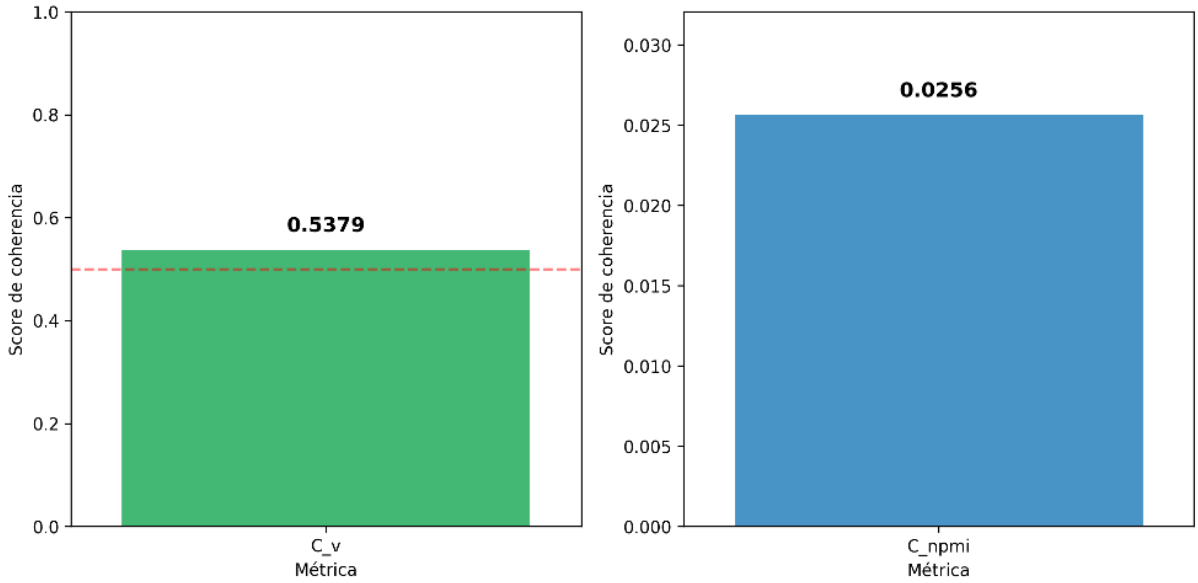
Nota. La figura muestra la evolución de la coherencia C_v del modelo LDA para distintas configuraciones del número de tópicos, evidenciando el valor máximo en k = 3.

Figura 17. Diversidad temática del modelo LDA según número de tópicos



Nota. La figura presenta la variación de la diversidad temática en función del número de tópicos evaluado, mostrando una disminución progresiva a medida que aumenta k.

Figura 18. Métricas finales de validación del modelo LDA seleccionado



Nota. La figura resume las métricas de coherencia del modelo LDA final seleccionado, incluyendo C_v y C_npmi.

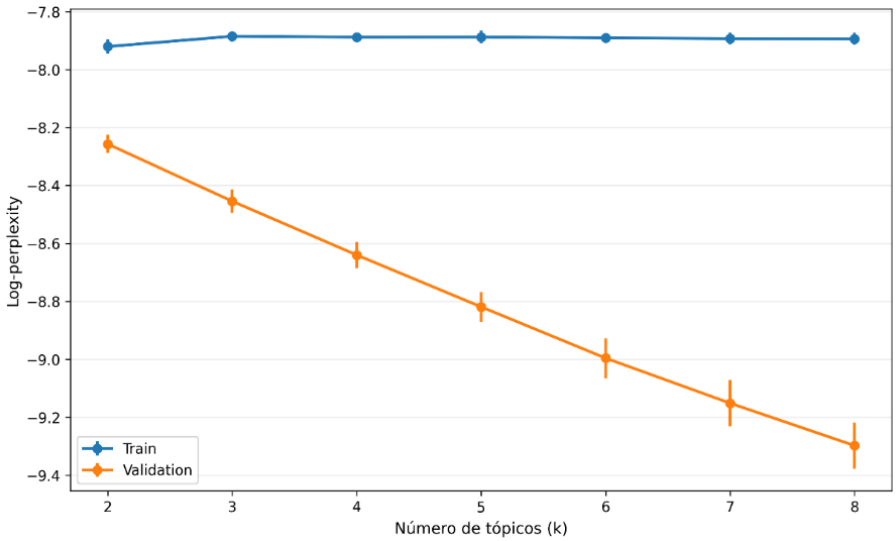
El valor C_v = 0.5379 corresponde al máximo observado dentro del rango de configuraciones evaluadas y, en ese sentido, respalda la selección relativa de k = 3 como solución óptima del modelo LDA. No obstante, este resultado no debe interpretarse como garantía de separación semántica perfecta ni como validación autosuficiente de las etiquetas temáticas. Su alcance es el de una evidencia de coherencia interna razonable dentro del espacio de búsqueda analizado, cuya robustez interpretativa se fortalece al considerarse conjuntamente con la coherencia C_npmi, la diversidad temática, el diagnóstico de ajuste, la estabilidad entre ejecuciones y la triangulación con modelos alternativos como HDP y BERTopic. En consecuencia, la interpretación de los tópicos se asume como una aproximación analítica plausible y metodológicamente sustentada, pero no como una estructura semántica cerrada o definitiva.

8.4.3. Diagnóstico de ajuste: underfitting y overfitting

Además de las métricas de coherencia semántica, el proceso de optimización incorporó un diagnóstico explícito del ajuste del modelo LDA, con el fin de identificar configuraciones potencialmente afectadas por subajuste (underfitting) o sobreajuste (overfitting). Para ello se comparó el comportamiento del modelo en conjuntos de entrenamiento y validación mediante log-perplejidad, así como el gap de generalización entre ambas curvas.

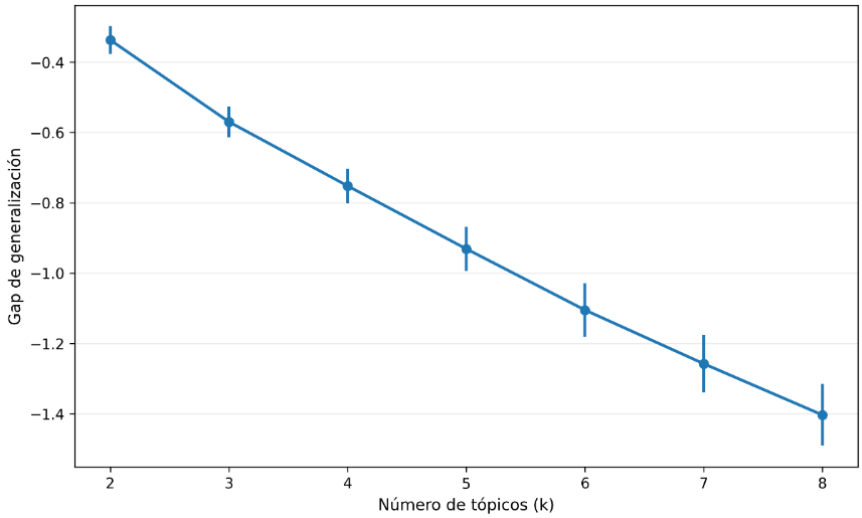
Los resultados indican que la log-perplejidad en entrenamiento se mantiene relativamente estable a medida que aumenta el número de tópicos, mientras que en validación mejora progresivamente (valores más bajos), generando una discrepancia creciente entre ambos conjuntos. Este patrón sugiere que configuraciones con valores altos de k capturan estructuras cada vez más específicas del corpus de entrenamiento, con menor capacidad de generalización sobre datos no observados.

Figura 19. Error de entrenamiento y validación del modelo LDA



Nota. La figura muestra la log-perplejidad del modelo LDA en entrenamiento y validación para distintos valores de k. La divergencia creciente entre ambas curvas se interpreta como evidencia de sobreajuste progresivo.

Figura 20. Gap de generalización del modelo LDA



Nota. La figura presenta la diferencia entre desempeño en entrenamiento y validación como indicador del riesgo de sobreajuste. Valores con mayor magnitud negativa reflejan mayor discrepancia entre ambos conjuntos.

El análisis del gap de generalización confirma esta tendencia: la diferencia entre entrenamiento y validación aumenta sistemáticamente en magnitud a medida que crece k, lo que constituye un indicador típico de sobreajuste en modelos probabilísticos. En contraste, valores bajos de k presentan menor discrepancia, aunque con riesgo de simplificación excesiva de la estructura temática.

En este contexto, la configuración con k = 3 representa un compromiso adecuado entre capacidad descriptiva y generalización, evitando tanto la fragmentación temática observada en configuraciones más granulares como la pérdida de información asociada a modelos excesivamente parsimoniosos.

8.4.4. Validación de estabilidad y robustez del modelo LDA

La selección del modelo final se complementó con un análisis de estabilidad estructural frente a diferentes inicializaciones del algoritmo. Dado que LDA es un modelo estocástico, los resultados pueden variar entre ejecuciones; por ello se evaluó la consistencia de las particiones temáticas mediante métricas de similitud inter-semilla.

La estabilidad se estimó con indicadores empleados en comparación de agrupamientos:

• NMI (Normalized Mutual Information)

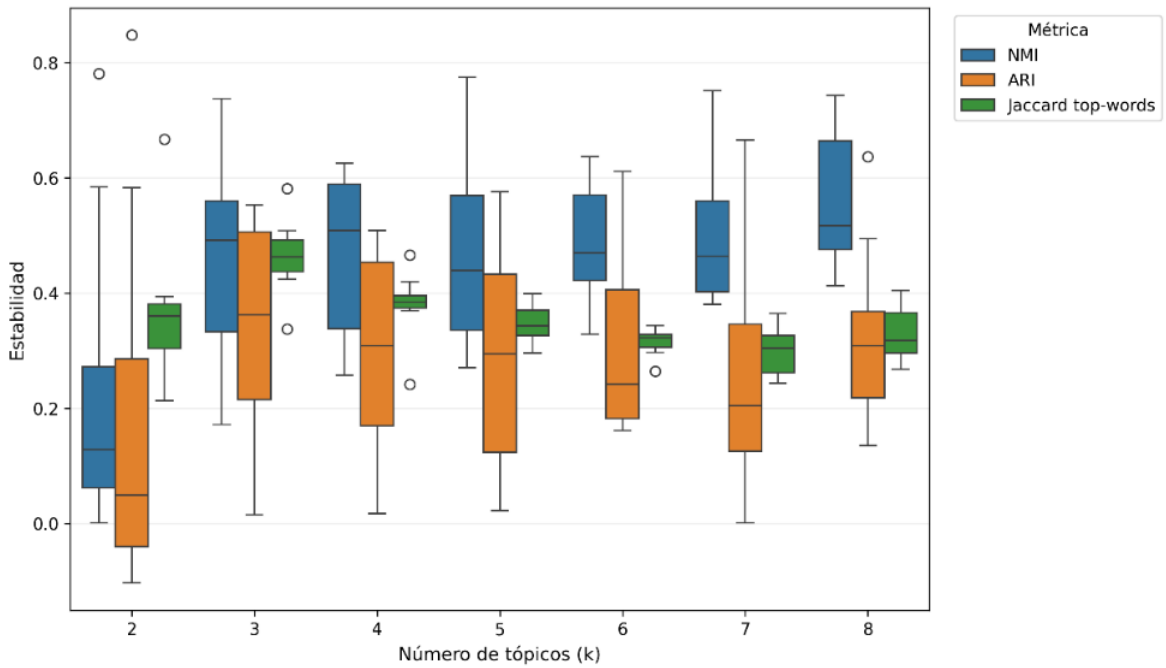
• ARI (Adjusted Rand Index)

• Jaccard sobre palabras dominantes de los tópicos

Los resultados muestran que, para valores intermedios de k, el modelo mantiene niveles moderados de consistencia estructural entre ejecuciones, mientras que configuraciones extremas presentan mayor variabilidad. En particular, el rango cercano a k = 3–5 exhibe una estabilidad adecuada sin sacrificar interpretabilidad.

Esta evidencia respalda que la solución final con k = 3 no depende críticamente de una inicialización específica y que los tópicos identificados corresponden a patrones persistentes del corpus. En términos metodológicos, esta robustez reduce la probabilidad de obtener estructuras temáticas espurias derivadas del azar algorítmico.

Figura 21. Estabilidad del modelo LDA según número de tópicos



Nota. La figura resume la estabilidad inter-ejecución del modelo LDA mediante NMI, ARI y Jaccard de términos dominantes. Valores más altos indican mayor consistencia estructural entre ejecuciones.

8.5. Triangulación con modelos alternativos no supervisados

La selección del modelo temático de referencia se sustentó en un proceso de triangulación metodológica que incorporó enfoques no supervisados con supuestos estadísticos distintos. Este procedimiento permitió contrastar la solución obtenida mediante LDA con modelos alternativos orientados a evaluar la plausibilidad estructural, la estabilidad y la coherencia de la organización temática identificada.

Se emplearon dos modelos complementarios: el Hierarchical Dirichlet Process (HDP), que infiere automáticamente el número de tópicos a partir de los datos, y BERTopic, basado en representaciones semánticas densas mediante embeddings contextuales. Mientras HDP se utilizó como referencia exploratoria para examinar la tendencia natural del corpus respecto al número de temas, BERTopic permitió un contraste estructural y semántico directo con el modelo LDA optimizado.

La triangulación resultante reduce la dependencia de un único método de modelado y aporta evidencia de que la estructura temática identificada refleja patrones consistentes del discurso mediático, y no un artefacto específico del algoritmo empleado.

8.5.1. Resultados del modelo HDP

El modelo HDP generó una estructura temática flexible sin necesidad de fijar previamente el número de tópicos. Los resultados muestran que la mayor parte de los documentos se concentra en un número reducido de temas activos, mientras que otros tópicos presentan contribuciones marginales.

Este comportamiento indica que el corpus tiende naturalmente hacia una organización temática compacta, lo que respalda la plausibilidad de una solución con pocos ejes narrativos dominantes. En este contexto, HDP no se utilizó como modelo competitivo frente a LDA, sino

como evidencia de apoyo para verificar que la selección de un número reducido de tópicos no impone artificialmente una estructura inexistente.

8.5.2. Resultados del modelo BERTopic

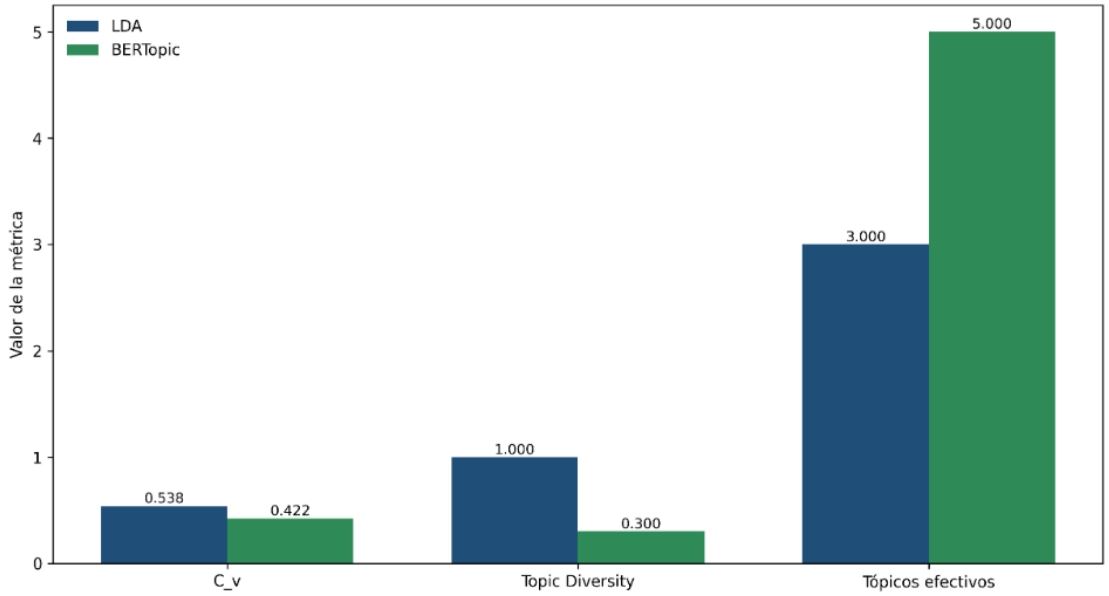
BERTopic identificó una estructura temática basada en similitud semántica contextual, con un mayor número de tópicos efectivos que LDA. Esta mayor granularidad refleja su capacidad para captar variaciones finas en el contenido discursivo, aunque introduce una fragmentación temática más elevada.

A diferencia de HDP, BERTopic permite una comparación directa con LDA en términos de coherencia, diversidad temática y asignación documental, por lo que se empleó como referencia principal para evaluar la robustez de la solución temática seleccionada.

8.5.3. Comparación cuantitativa entre modelos

La comparación muestra que LDA presenta mayor coherencia y máxima diversidad temática, lo que indica una separación clara entre tópicos y una estructura interpretativamente sólida. BERTopic, por su parte, identifica un mayor número de tópicos efectivos, lo que sugiere una representación más granular del corpus, aunque con menor cohesión interna.

Figura 22. Comparación de métricas temáticas entre LDA y BERTopic

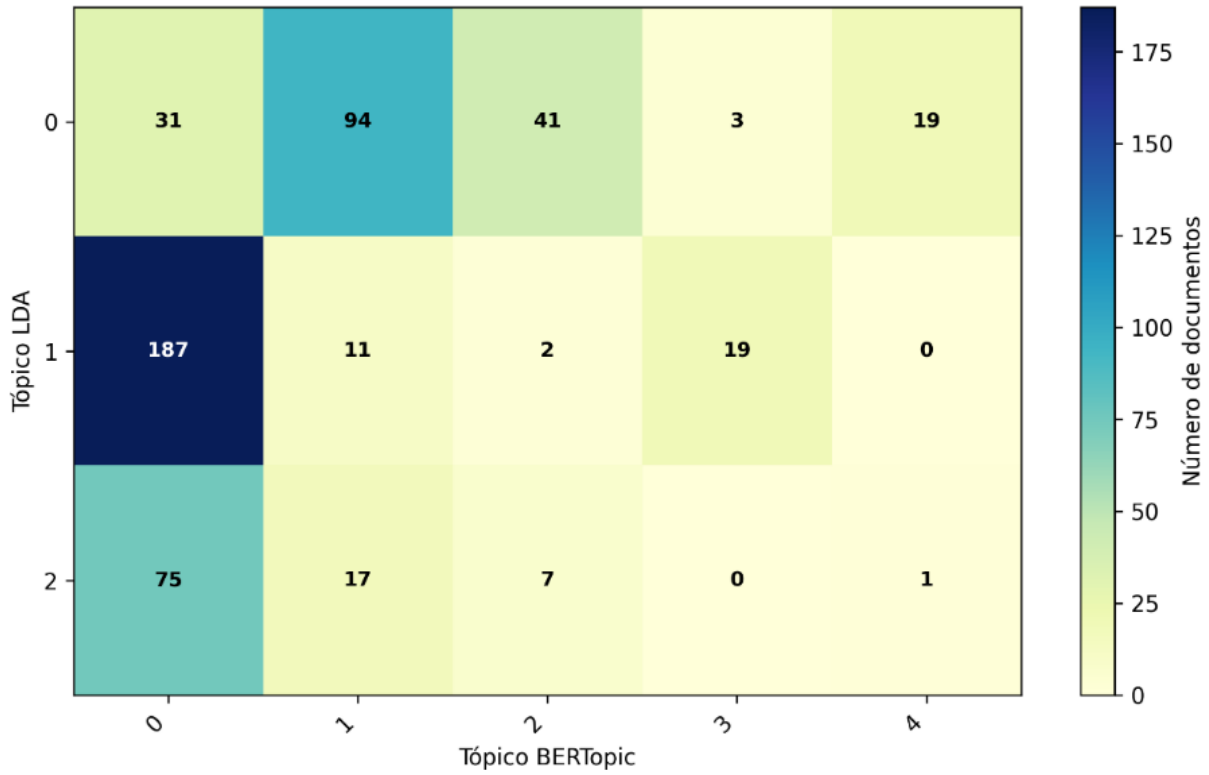


Nota. La figura compara coherencia semántica (C_v), diversidad temática y número de tópicos efectivos entre LDA y BERTopic, utilizando métricas equivalentes para evaluar la consistencia estructural de ambos modelos.

8.5.4. Consistencia intermodelo y alineación temática

La correspondencia evidencia que los tópicos LDA concentran documentos distribuidos en múltiples tópicos de BERTopic, lo que indica que este último subdivide ejes narrativos amplios en subtemas más específicos.

Figura 23. Correspondencia documental entre tópicos LDA y BERTopic



Nota. La matriz de contingencia muestra la distribución de documentos asignados a cada combinación de tópicos entre ambos modelos, permitiendo identificar solapamientos y subdivisiones temáticas.

La matriz incluye únicamente los documentos con asignación temática válida en ambos modelos. Los registros clasificados como ruido por BERTopic o sin correspondencia directa fueron excluidos del análisis comparativo.

8.5.5. Selección final del modelo temático de referencia

Considerando conjuntamente los resultados de coherencia, diversidad, granularidad y consistencia intermodelo, el modelo LDA optimizado se adoptó como referencia principal para el análisis posterior del discurso mediático.

Esta decisión se fundamenta en que LDA proporciona una estructura temática compacta, interpretable y coherente, al tiempo que mantiene correspondencia sustancial con las estructuras identificadas por los modelos alternativos. La triangulación confirma así que la solución de tres tópicos representa una organización estable del espacio temático del corpus.

En consecuencia, los análisis posteriores se desarrollan utilizando LDA como modelo base, incorporando la evidencia aportada por HDP y BERTopic únicamente como respaldo metodológico.

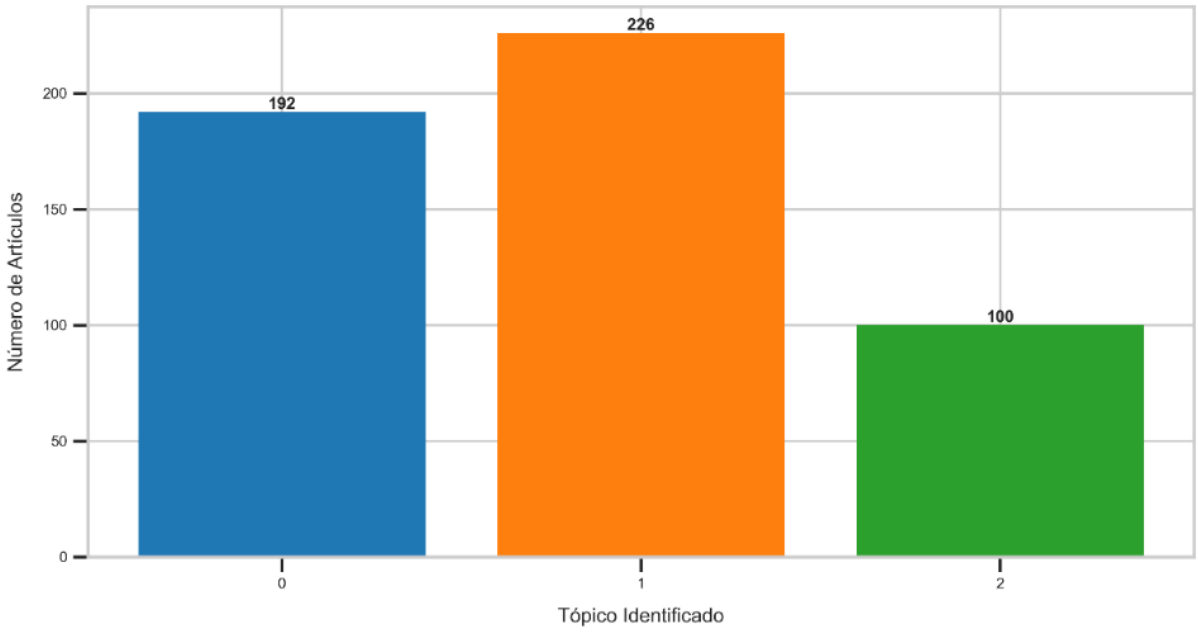
8.6. Estructura temática del discurso mediático (modelo seleccionado)

Una vez validado el modelo LDA como referencia principal, se analizó la estructura temática latente del corpus con el fin de identificar los ejes narrativos dominantes del discurso mediático sobre corrupción en el sector salud. El modelo optimizado (k = 3) permite representar el espacio discursivo mediante tres tópicos principales que concentran la totalidad de los documentos analizados.

Esta configuración sugiere que el fenómeno es abordado mediáticamente a través de un número limitado de marcos temáticos recurrentes, cada uno asociado a vocabularios específicos y patrones narrativos diferenciados. Los apartados siguientes describen la distribución, contenido semántico y comportamiento contextual de estos ejes.

8.6.1. Distribución global de los ejes temáticos

Figura 24. Distribución global de documentos por tópico del modelo LDA



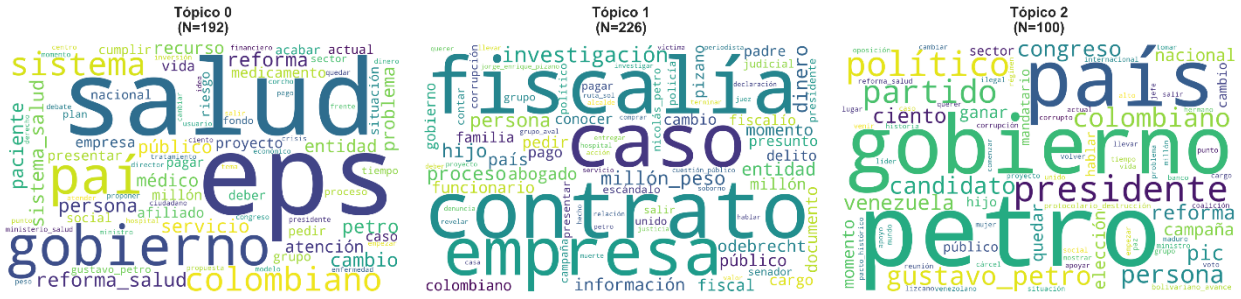
Nota. La figura muestra la proporción de documentos asignados a cada tópico del modelo LDA optimizado, evidenciando el peso relativo de los ejes temáticos dentro del corpus.

Los resultados indican una distribución desigual entre los tópicos, con un eje dominante que concentra la mayor proporción de artículos, seguido por dos ejes secundarios con menor presencia relativa. Esta asimetría sugiere que ciertos marcos interpretativos ocupan un lugar central en la cobertura mediática, mientras que otros representan dimensiones más específicas del fenómeno.

8.6.2. Interpretación semántica de los tópicos

El análisis léxico muestra que cada tópico está asociado a conjuntos de términos distintivos, lo que confirma la separación semántica entre los ejes narrativos. Estas diferencias permiten interpretar los tópicos como representaciones de dimensiones específicas del discurso sobre corrupción en salud, vinculadas a actores, prácticas institucionales y dinámicas del sistema.

Figura 25. Nubes léxicas de los tópicos del modelo LDA



Nota. La figura sintetiza las palabras con mayor peso probabilístico en cada tópico, permitiendo una interpretación cualitativa de los ejes temáticos identificados.

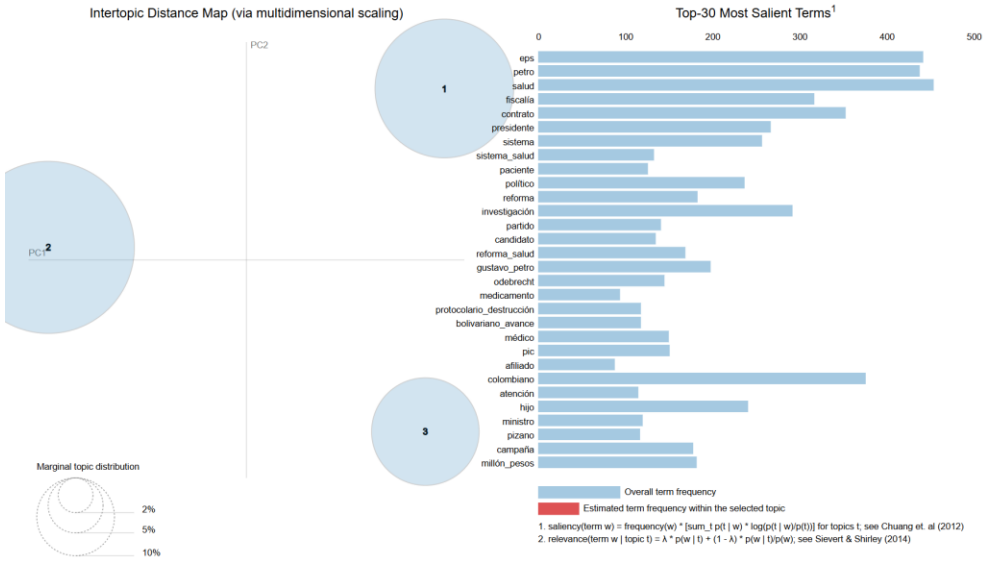
Las etiquetas temáticas propuestas deben entenderse como interpretaciones guiadas por evidencia probabilística y validación comparativa, sujetas a los límites de coherencia y separación semántica propios del modelo seleccionado.

El proceso de asignación de etiquetas conceptuales a los tópicos fue realizado por un único investigador, a partir de la revisión de los términos con mayor peso probabilístico y de los documentos más representativos asociados a cada tópico, contrastados con el marco teórico del estudio y con la tipología de marcos narrativos definida previamente. No se implementó un procedimiento formal de validación inter-jueces ni revisión experta externa para esta fase interpretativa. En consecuencia, la trazabilidad del etiquetado se apoyó en criterios explícitos de interpretación, en la consistencia entre términos dominantes y documentos representativos, y en mecanismos indirectos de control analítico, como la coherencia semántica, la diversidad temática, la estabilidad del modelo y la triangulación con HDP y BERTopic.

8.6.3. Composición léxica y separación inter-tópica

La separación observable entre los círculos temáticos indica que los tópicos presentan solapamientos limitados y una diferenciación estructural clara. Este patrón respalda la coherencia del modelo y su capacidad para capturar dimensiones discursivas distintas sin fragmentación excesiva.

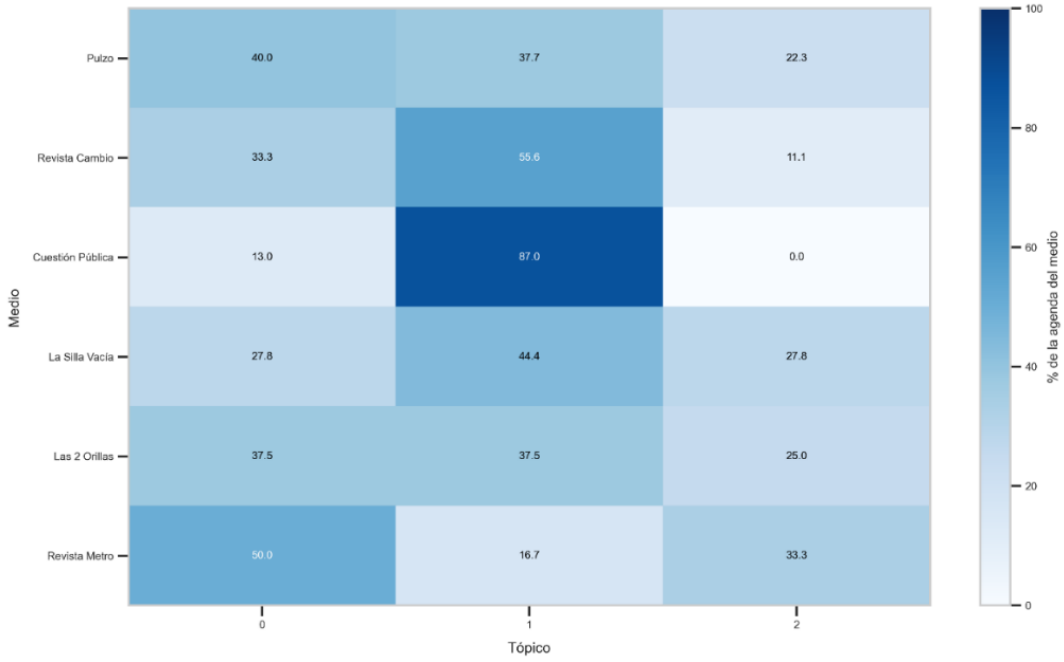
Figura 26. Visualización LDAvis de la separación inter-tópica



Nota. La visualización muestra la distancia semántica entre tópicos en el espacio de términos, así como la relevancia de las palabras más representativas de cada uno.

8.6.4. Variación temática por medio digital

Figura 27. Distribución de tópicos por medio de comunicación



Nota. El mapa de calor muestra la intensidad relativa de cada tópico según el medio digital, permitiendo identificar diferencias en la agenda temática editorial.

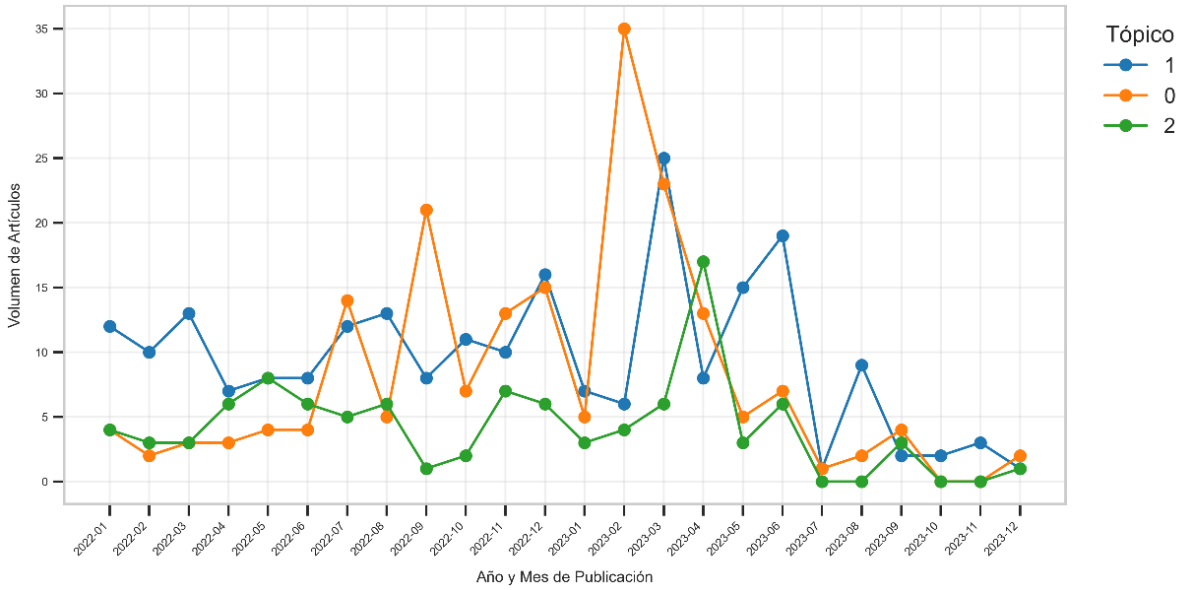
El mapa de calor muestra que la distribución temática varía entre los medios digitales analizados, evidenciando diferencias en la agenda editorial. El Tópico 1 presenta una alta concentración en Cuestión Pública, mientras que medios como Pulzo, La Silla Vacía y Las 2 Orillas exhiben una distribución más equilibrada entre los tres tópicos. Por su parte, Revista Metro muestra predominio del Tópico 0, y Revista Cambio concentra una mayor proporción en el Tópico 1, aunque con presencia de los demás ejes temáticos.

En conjunto, estos resultados indican que la cobertura del fenómeno no es homogénea entre medios, sino que cada uno prioriza distintos contenidos dentro del mismo campo informativo.

8.6.5. Evolución temporal de la agenda temática

La dinámica temporal indica fluctuaciones en la prominencia de los ejes temáticos a lo largo del periodo analizado. No obstante, los tres tópicos se mantienen presentes durante todo el intervalo temporal, lo que sugiere la persistencia de marcos estructurales en la cobertura mediática.

Figura 28. Evolución temporal de los tópicos LDA



Nota. La figura muestra la variación del peso de cada tópico a lo largo del periodo analizado, permitiendo identificar cambios en la agenda mediática.

8.7. Análisis de sentimiento y polarización discursiva

El análisis de sentimiento se orientó a identificar el tono emocional predominante en la cobertura mediática sobre corrupción en el sector salud, así como los patrones de polarización asociados a los distintos ejes temáticos, medios y periodos temporales. Para ello se utilizó un enfoque multifuente que combina un modelo supervisado basado en aprendizaje profundo con recursos léxicos especializados en español.

Este procedimiento permite capturar tanto evaluaciones contextuales complejas como tendencias generales del lenguaje emocional, reduciendo las limitaciones inherentes a un único método de medición. Los resultados se presentan a continuación desde una perspectiva estructural, considerando la confiabilidad del modelo contextual, la comparación entre métodos y la distribución de la negatividad en relación con tópicos, medios y tiempo.

8.7.1. Enfoque metodológico multifuente

El análisis principal se realizó mediante un modelo RoBERTa entrenado para clasificación de sentimiento en español, complementado con dos baselines léxicos: ML-SentiCon y NRC EmoLex. Esta combinación permite contrastar evaluaciones basadas en contexto semántico profundo con aproximaciones basadas en polaridad léxica.

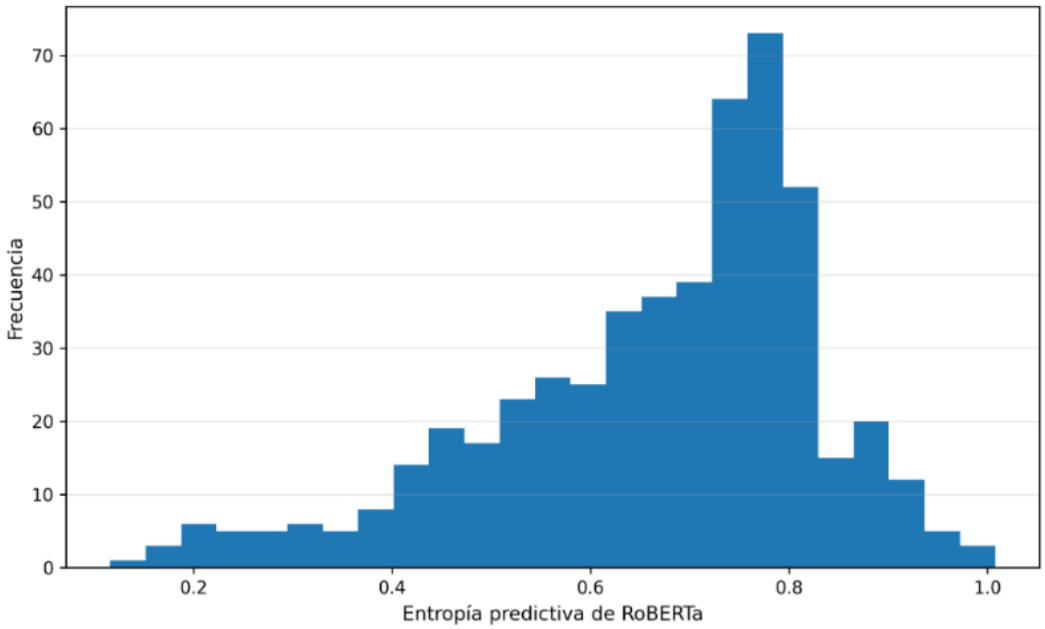
El modelo supervisado aporta mayor sensibilidad a fenómenos discursivos como ironía, atribución indirecta o construcción narrativa, mientras que los recursos léxicos permiten validar tendencias generales y detectar posibles sesgos metodológicos. Esta triangulación es consistente con los criterios metodológicos planteados en el marco teórico, particularmente en relación con la comparación documental entre métodos y la evaluación de consistencia de resultados.

8.7.2. Validación y confiabilidad del análisis de sentimiento

La confiabilidad del modelo contextual se evaluó mediante dos indicadores complementarios: la entropía predictiva y el margen de confianza. Ambos permiten estimar el grado de incertidumbre de las clasificaciones generadas por RoBERTa.

La distribución observada indica que una proporción importante de documentos se concentra en niveles medios y altos de entropía, lo que sugiere que el corpus contiene textos con carga semántica compleja y con componentes de ambigüedad emocional. Este comportamiento es consistente con el carácter periodístico del corpus, donde la narración factual, la cita indirecta y el lenguaje institucional pueden reducir la separación entre clases.

Figura 29. Distribución de entropía del modelo RoBERTa

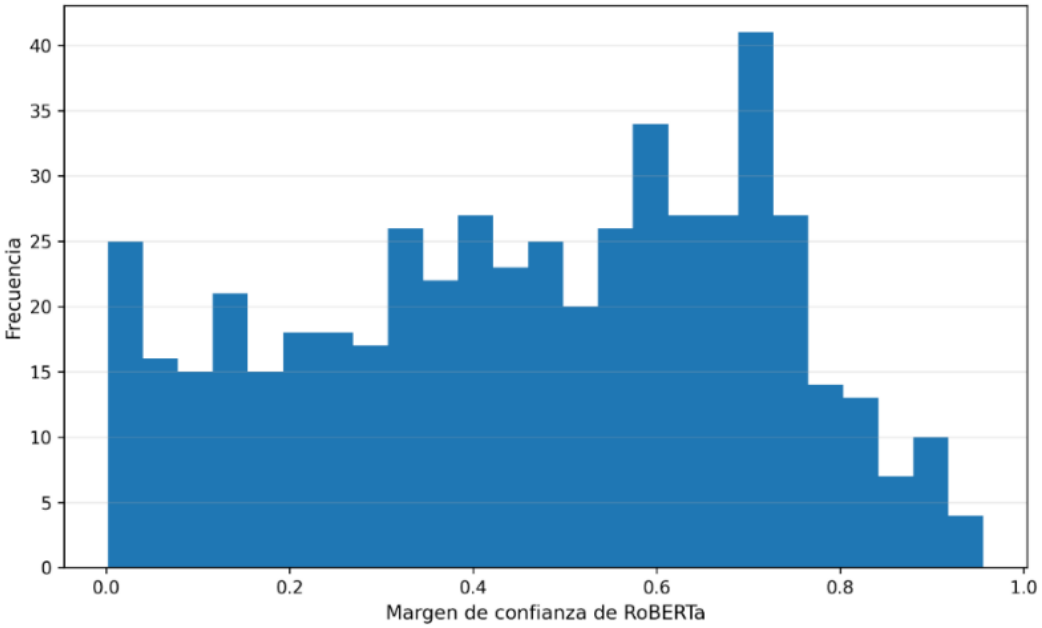


Nota. La figura muestra la distribución de la entropía predictiva de RoBERTa en el corpus analizado. Valores más altos indican mayor incertidumbre en la asignación de polaridad.

Complementariamente, la distribución del margen de confianza muestra una concentración importante de documentos en valores intermedios y altos, lo que indica que, a nivel agregado, el modelo conserva capacidad discriminativa suficiente para sustentar el análisis posterior. En conjunto, ambos indicadores sugieren que las predicciones de RoBERTa

son utilizables para fines analíticos, aunque reflejan la complejidad inherente al dominio discursivo estudiado.

Figura 30. Margen de confianza en la clasificación de sentimiento



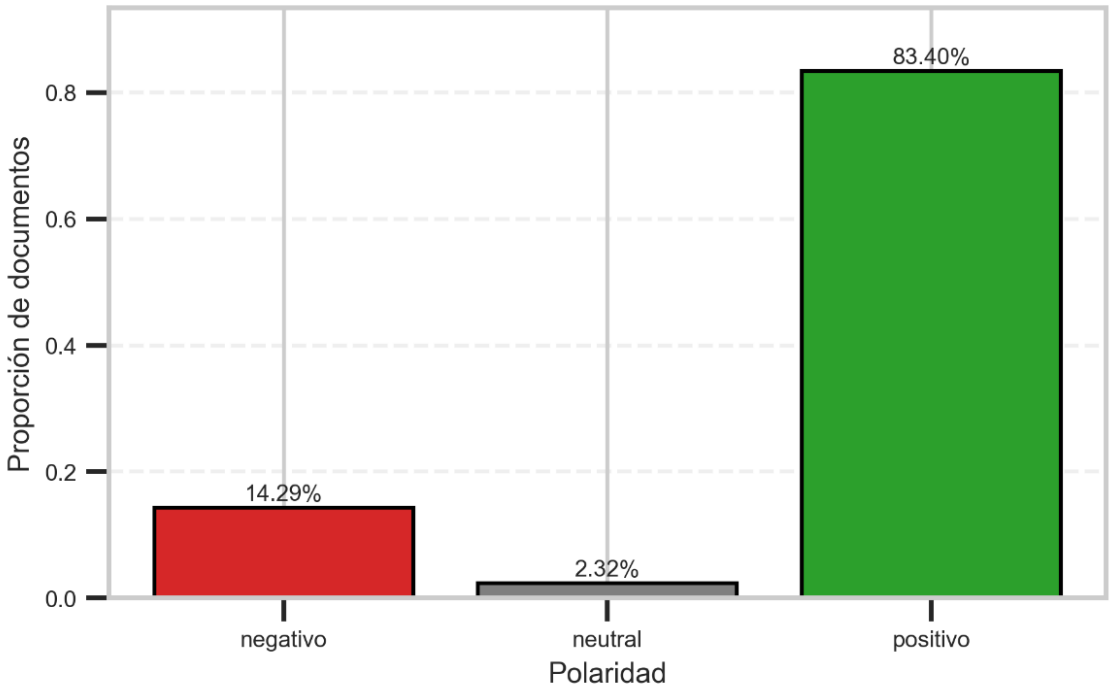
Nota. La figura presenta la distribución del margen de confianza de RoBERTa, calculado como la diferencia entre las dos probabilidades de clase más altas para cada documento.

8.7.3. Consistencia entre métodos de medición emocional

Como punto de contraste, se examinaron las distribuciones globales de polaridad obtenidas mediante los recursos léxicos ML-SentiCon y NRC EmoLex. Ambos métodos muestran predominancia de polaridad positiva, aunque con distinta intensidad relativa.

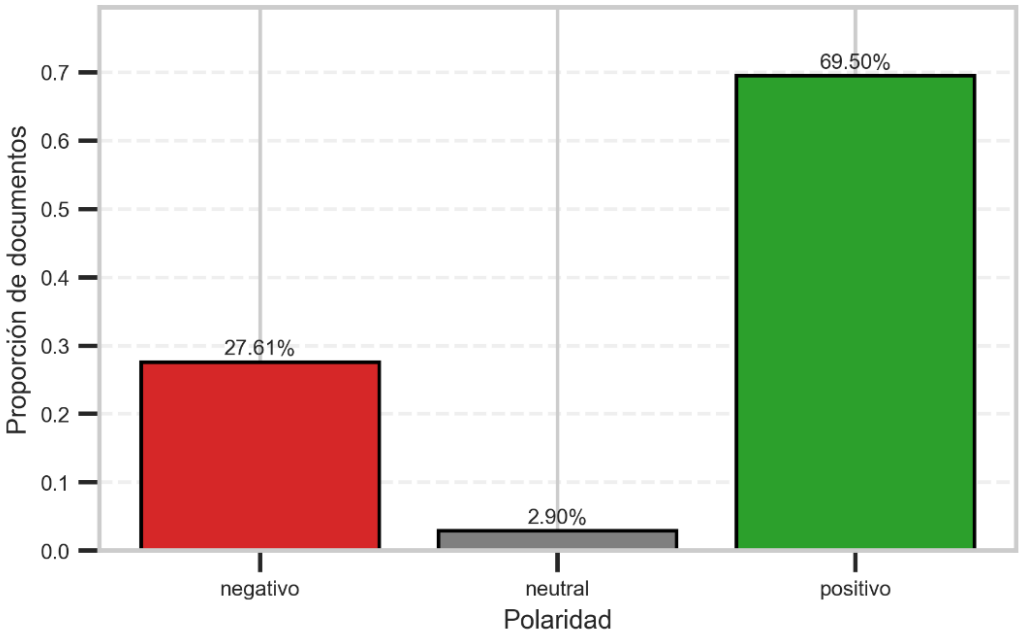
La coincidencia entre ambos enfoques léxicos sugiere estabilidad en la orientación afectiva superficial del vocabulario empleado en el corpus. Sin embargo, estos resultados no deben interpretarse como evidencia definitiva del tono discursivo, ya que los métodos léxicos operan sobre asociaciones palabra-polaridad y no incorporan el contexto completo de enunciación. En textos periodísticos sobre corrupción, esta limitación puede conducir a una sobreestimación de polaridad positiva o neutra cuando predominan términos institucionales, descriptivos o procedimentales.

Figura 31. Polaridad léxica del corpus según ML-SentiCon



Nota. La figura presenta la distribución global de polaridad obtenida mediante ML-SentiCon en las categorías negativo, neutral y positivo.

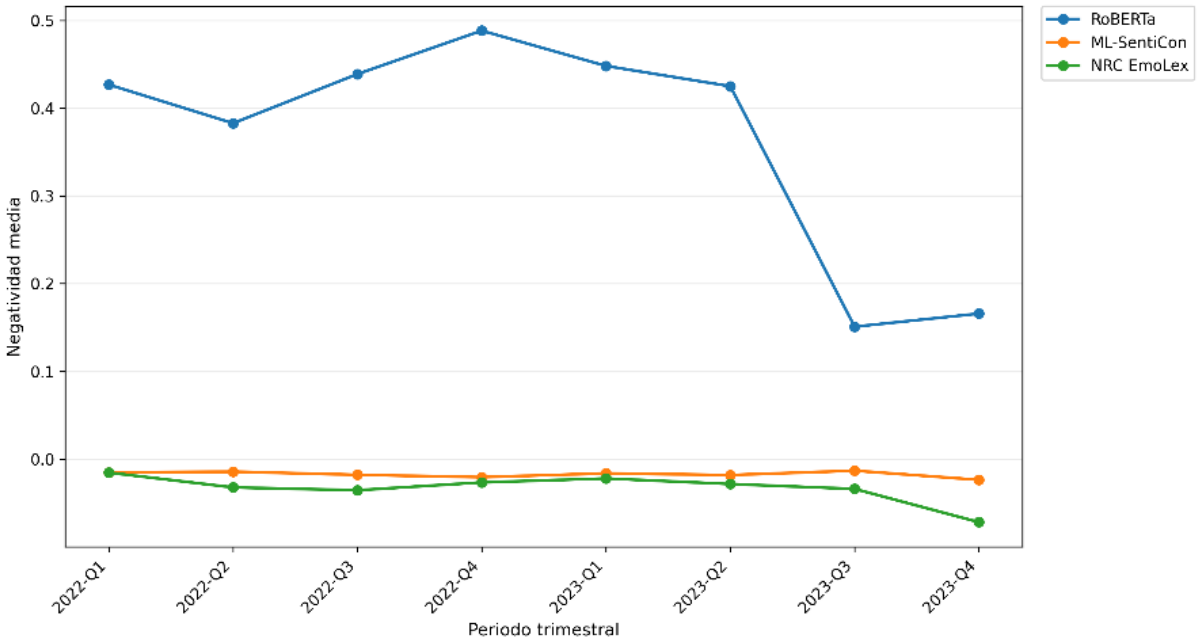
Figura 32. Polaridad léxica del corpus según NRC EmoLex



Nota. La figura presenta la distribución global de polaridad obtenida mediante NRC EmoLex en las categorías negativo, neutral y positivo.

Para evaluar de forma más directa la relación entre métodos, se comparó la evolución temporal de la negatividad estimada por RoBERTa, ML-SentiCon y NRC EmoLex.

Figura 33. Evolución temporal comparativa de la negatividad entre métodos



Nota. La figura compara la negatividad media trimestral estimada por RoBERTa, ML-SentiCon y NRC EmoLex durante el periodo 2022–2023.

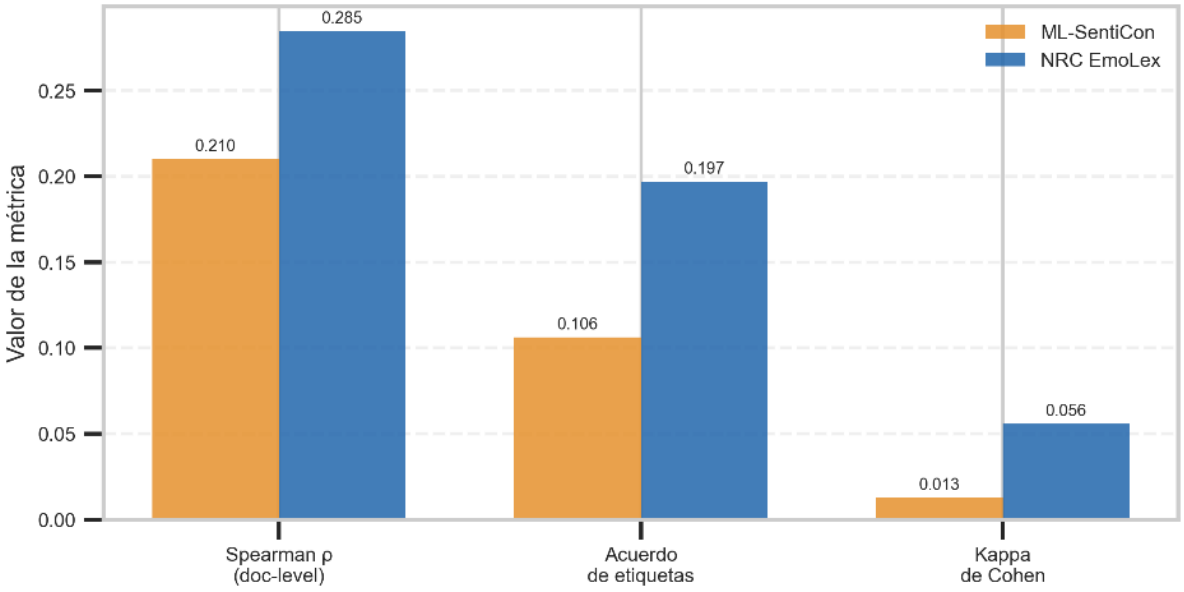
La serie evidencia una divergencia sistemática entre el modelo contextual y los enfoques léxicos. Mientras RoBERTa registra niveles de negatividad considerablemente más altos, los dos recursos léxicos se mantienen en valores próximos a cero y con variaciones mucho más reducidas. Este patrón refuerza la idea de que la polaridad léxica no capta adecuadamente la carga crítica del discurso periodístico sobre corrupción, la cual depende en mayor medida de relaciones semánticas y contextuales.

Con el fin de sintetizar cuantitativamente esta comparación, se calcularon métricas de correlación documental, acuerdo de etiquetas y consistencia categórica entre los métodos.

Los resultados muestran que NRC EmoLex presenta valores superiores a ML-SentiCon en las tres métricas comparadas, aunque en ambos casos el nivel de acuerdo sigue siendo limitado. En términos metodológicos, esto indica que los recursos léxicos conservan utilidad

como referencia de contraste, pero no sustituyen la capacidad contextual del modelo supervisado. Esta evidencia permite respaldar explícitamente los criterios expuestos en el marco teórico sobre correlación documental y consistencia entre métodos.

Figura 34. Panel comparativo de consistencia entre métodos de análisis de sentimiento



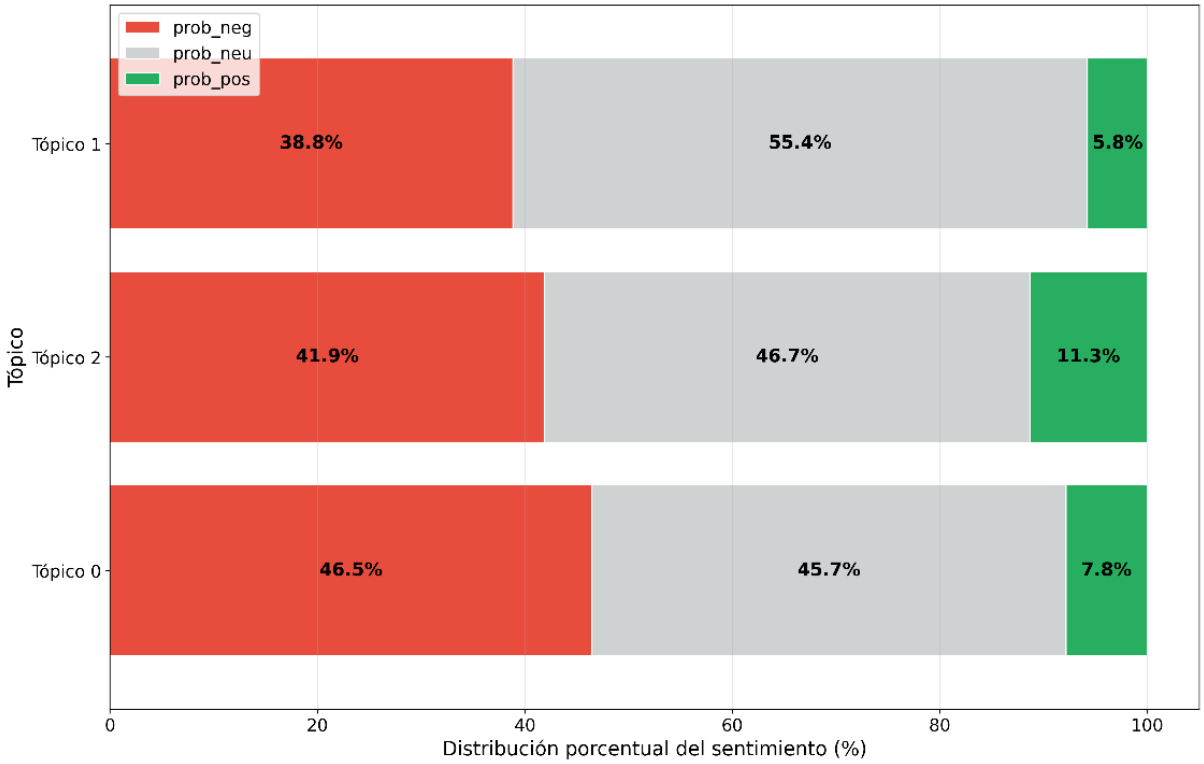
Nota. La figura presenta la consistencia relativa de los recursos léxicos ML-SentiCon y NRC EmoLex a nivel documental y categórico, evaluada mediante correlación de Spearman, acuerdo de etiquetas y kappa de Cohen.

En consecuencia, el modelo contextual se mantiene como referencia principal no porque carezca de incertidumbre, sino porque conserva capacidad discriminativa suficiente y supera a los enfoques léxicos en la captura de relaciones semánticas y contextuales del discurso periodístico analizado.

8.7.4. Polarización negativa por tópico y medio

A diferencia de los enfoques léxicos, el modelo contextualizado revela un predominio claro de negatividad cuando se examina la distribución del sentimiento en relación con la estructura temática del corpus.

Figura 35. Distribución del sentimiento por eje temático

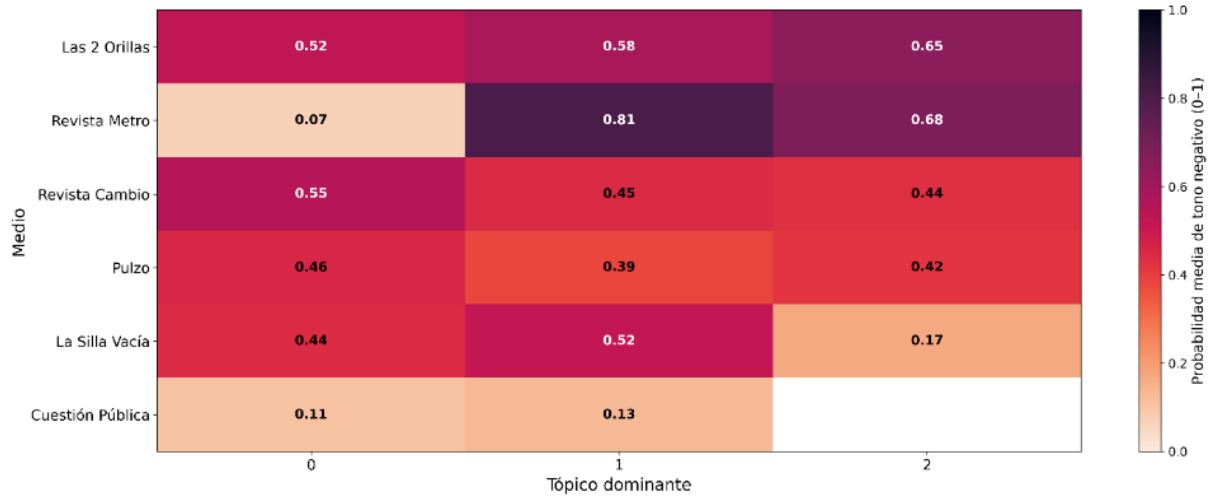


Nota. La figura muestra la proporción porcentual de probabilidad negativa, neutral y positiva en cada tópico dominante del modelo LDA.

Los resultados evidencian que los tres tópicos presentan una presencia relevante de negatividad, aunque con intensidades diferentes. Esto sugiere que el tono crítico no se concentra en un único eje narrativo, sino que constituye un rasgo transversal del discurso mediático sobre corrupción en salud. La variación observada entre tópicos indica, además, que la carga emocional se distribuye de manera desigual según el contenido temático predominante.

Con el fin de refinar esta lectura, se examinó la intensidad de negatividad según la combinación entre medio digital y tópico dominante.

Figura 36. Intensidad de tono negativo por medio y tópico



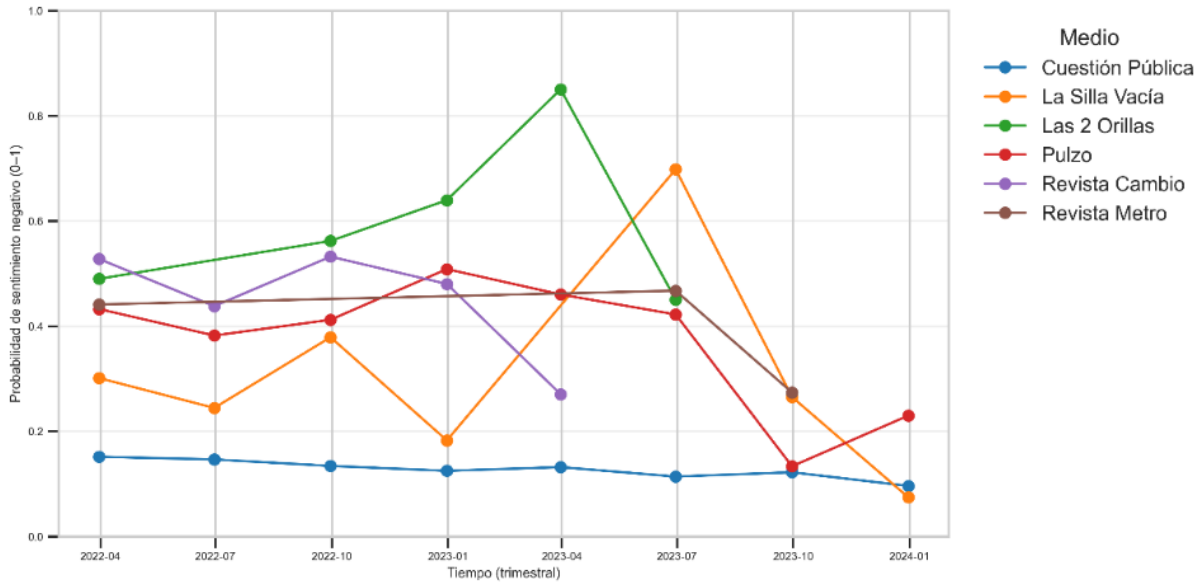
Nota. El mapa de calor muestra la probabilidad media de tono negativo para cada combinación de medio digital y tópico dominante. Las celdas en blanco indican ausencia de documentos en esa combinación.

La figura permite identificar diferencias sustantivas entre medios en la intensidad del tono negativo, así como variaciones internas según el eje temático cubierto. Algunos medios exhiben niveles elevados de negatividad en varios tópicos, mientras que otros muestran patrones más acotados o dependientes del contenido específico. En consecuencia, la polarización negativa no solo depende del tema tratado, sino también del estilo editorial con que cada medio construye discursivamente la corrupción en salud.

8.7.5. Evolución temporal del tono emocional

Finalmente, se analizó la evolución temporal de la negatividad discursiva por medio digital, con el fin de identificar patrones de estabilidad y variación durante el periodo de estudio.

Figura 37. Evolución temporal de la negatividad discursiva por medio



Nota. La figura presenta la evolución trimestral de la probabilidad media de sentimiento negativo en cada medio digital incluido en el corpus.

La dinámica temporal muestra fluctuaciones en la intensidad del tono negativo entre medios y periodos, aunque se mantiene una presencia sostenida de negatividad a lo largo de todo el intervalo analizado. Este comportamiento sugiere que el tratamiento discursivo de la corrupción en salud conserva un componente crítico estructural, aun cuando su intensidad varía según el contexto temporal y la agenda editorial de cada fuente.

8.7.6. Resultados formales de la validación estadística no paramétrica

A continuación, se presentan los resultados de la validación estadística no paramétrica aplicada a las diferencias en la distribución del tono emocional entre grupos discursivos. En coherencia con el diseño metodológico, se empleó la prueba de Kruskal-Wallis para evaluar heterogeneidad global entre grupos independientes y, cuando correspondió, contrastes post-hoc mediante Mann-Whitney U con corrección de Holm. Como medida de magnitud del efecto se reporta epsilon cuadrado (ε^2) para los contrastes globales y Cliff's Delta (δ) para las comparaciones pareadas.

Tabla 19. Resultados globales de la validación estadística no paramétrica

Variable / contraste	Agrupación	Prueba	Estadístico	Valor p	Tamaño del efecto	Interpretación
prob_neg	Medio	Kruskal-Wallis	H = 0.6094	0.4350	$\varepsilon^2 = 0.0000$	No se evidencian diferencias globales significativas
prob_neg	Tópico	Kruskal-Wallis	H = 8.6613	0.0132	$\varepsilon^2 = 0.0925$	Se evidencian diferencias globales significativas
acus_density_x1000	Medio	Kruskal-Wallis	H = 16.4697	0.000049	$\varepsilon^2 = 0.2344$	Diferencias globales significativas
acus_density_x1000	Tópico	Kruskal-Wallis	H = 6.4977	0.0388	$\varepsilon^2 = 0.0625$	Diferencias globales significativas
score_doc	Medio	Kruskal-Wallis	H = 16.7420	0.000043	$\varepsilon^2 = 0.2385$	Diferencias globales significativas
score_doc	Tópico	Kruskal-Wallis	H = 5.0292	0.0809	$\varepsilon^2 = 0.0421$	No se evidencian diferencias globales significativas

Nota. La tabla presenta los resultados globales de la validación estadística no paramétrica aplicada a variables documentales derivadas del corpus. Se utilizó la prueba de Kruskal-Wallis para evaluar heterogeneidad entre grupos independientes, reportando el estadístico H, el valor p y el tamaño del efecto mediante epsilon cuadrado (ε^2).

En relación directa con la variable principal de tono emocional, operacionalizada mediante la probabilidad de polaridad negativa (prob_neg), no se observaron diferencias estadísticamente significativas entre medios ($H = 0.6094$; $p = 0.4350$; $\varepsilon^2 = 0.0000$), lo que indica ausencia de heterogeneidad global relevante en esta dimensión entre los grupos que cumplieron el umbral mínimo de observaciones para la comparación. En cambio, sí se identificaron diferencias significativas por tópico ($H = 8.6613$; $p = 0.0132$; $\varepsilon^2 = 0.0925$), con un tamaño del efecto pequeño a moderado.

El análisis post-hoc para prob_neg por tópico mostró que la única diferencia estadísticamente significativa, tras corrección de Holm, se presenta entre los tópicos 0 y 1 ($U = 538$; p ajustado = 0.0072; $\delta = 0.4841$), mientras que los contrastes entre 0 y 2, y entre 1 y 2, no alcanzaron significación estadística. En términos sustantivos, estos resultados respaldan que la variación del tono emocional del corpus no se distribuye de manera homogénea entre los ejes

temáticos identificados, aunque dicha variación no se reproduce con la misma intensidad entre medios.

Dado que la tabla sintetiza únicamente los contrastes post-hoc con significancia estadística, no se reportan todas las comparaciones pareadas entre grupos, sino solo aquellas que mantuvieron significancia tras la corrección por comparaciones múltiples.

Tabla 20. Contrastes post-hoc estadísticamente significativos validación no paramétrica

Variable / contraste	Comparación	Prueba	Estadístico	Valor p ajustado (Holm)	Tamaño del efecto	Interpretación
prob_neg por tópico	Tópico 0 vs Tópico 1	Mann-Whitney U	U = 538	0.0072	δ = 0.4841	Diferencia significativa, efecto moderado
acus_density_x1000 por medio	Pulzo vs Revista Cambio	Mann-Whitney U	U = 781.5	0.000051	δ = 0.6281	Diferencia significativa, efecto moderado-alto
score_doc por medio	Pulzo vs Revista Cambio	Mann-Whitney U	U = 784	0.000044	δ = 0.6333	Diferencia significativa, efecto moderado-alto

Nota. La tabla presenta únicamente los contrastes post-hoc que conservaron significancia estadística tras la prueba global de Kruskal-Wallis y la corrección de Holm para comparaciones múltiples. Las comparaciones pareadas se realizaron mediante Mann-Whitney U y se reporta el tamaño del efecto mediante Cliff's Delta (δ). Los contrastes no significativos no se incluyen en esta tabla por razones de síntesis expositiva.

De forma complementaria, la validación no paramétrica mostró diferencias significativas en métricas de atribución discursiva como acus_density_x1000 y score_doc, tanto por medio como por tópico en algunos contrastes, lo que aporta evidencia adicional sobre la heterogeneidad del corpus más allá del componente estrictamente emocional. No obstante, dado que la contrastación principal de H_0 y H_1 se relaciona con el tono emocional del discurso, el énfasis interpretativo de esta subsección recae en los resultados asociados a prob_neg.

En conjunto, estos resultados permiten sustentar formalmente la contrastación de las hipótesis de la investigación en el componente emocional del discurso. En particular, la evidencia estadística indica que las diferencias observadas no se limitan a variaciones

descriptivas, sino que presentan soporte inferencial en los contrastes no paramétricos aplicados, particularmente en los contrastes asociados a los ejes temáticos identificados. No obstante, estas pruebas deben interpretarse en articulación con el análisis temático y con la triangulación metodológica general del estudio, dado que la robustez de los hallazgos no depende de una sola métrica ni de un único componente analítico.

8.8. Marcos narrativos y atribución de responsabilidad

El análisis de marcos narrativos (frames) se orientó a identificar las estructuras interpretativas mediante las cuales los medios construyen discursivamente la corrupción en el sector salud. A diferencia del modelado temático, que describe el contenido semántico predominante, el análisis de frames permite examinar la forma en que los hechos son presentados, qué dimensiones se enfatizan y a qué actores se atribuye responsabilidad.

Este componente integra resultados del modelado temático, el análisis de sentimiento y la identificación de actores, con el fin de reconstruir patrones narrativos recurrentes en la cobertura mediática. Los apartados siguientes describen los marcos dominantes, su distribución entre medios, su evolución temporal y los procesos de atribución discursiva.

En este estudio, los marcos narrativos se operacionalizan mediante un enfoque léxico-computacional basado en la densidad de palabras clave asociadas a cada categoría narrativa dentro de los documentos analizados.

Los marcos identificados corresponden a patrones léxicos vinculados a determinadas narrativas, no a marcos discursivos en sentido hermenéutico estricto. En consecuencia, los resultados deben interpretarse como indicadores cuantitativos de resonancia temática y no como reconstrucciones completas de estructuras narrativas o interpretativas.

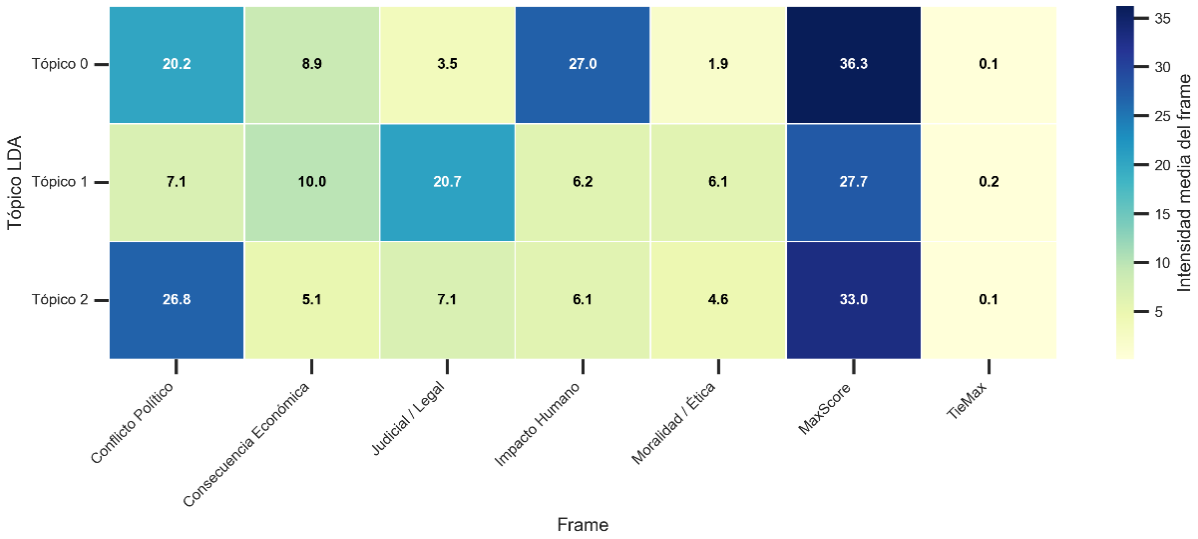
8.8.1. Identificación de frames dominantes

Los resultados evidencian una asociación diferenciada entre los tópicos del modelo LDA y los marcos narrativos identificados. El marco de conflicto político presenta mayor intensidad

en el Tópico 2 (26.8), mientras que el Tópico 1 se caracteriza por una presencia destacada del encuadre judicial/legal (20.7). El Tópico 0 combina niveles relevantes de conflicto político (20.2) e impacto humano (27.0).

El indicador MaxScore refuerza esta diferenciación al registrar valores elevados en los tres tópicos (36.3 en Tópico 0, 27.7 en Tópico 1 y 33.0 en Tópico 2), lo que sugiere que cada eje temático tiende a estructurarse alrededor de un encuadre predominante.

Figura 38. Distribución de marcos narrativos por tópico



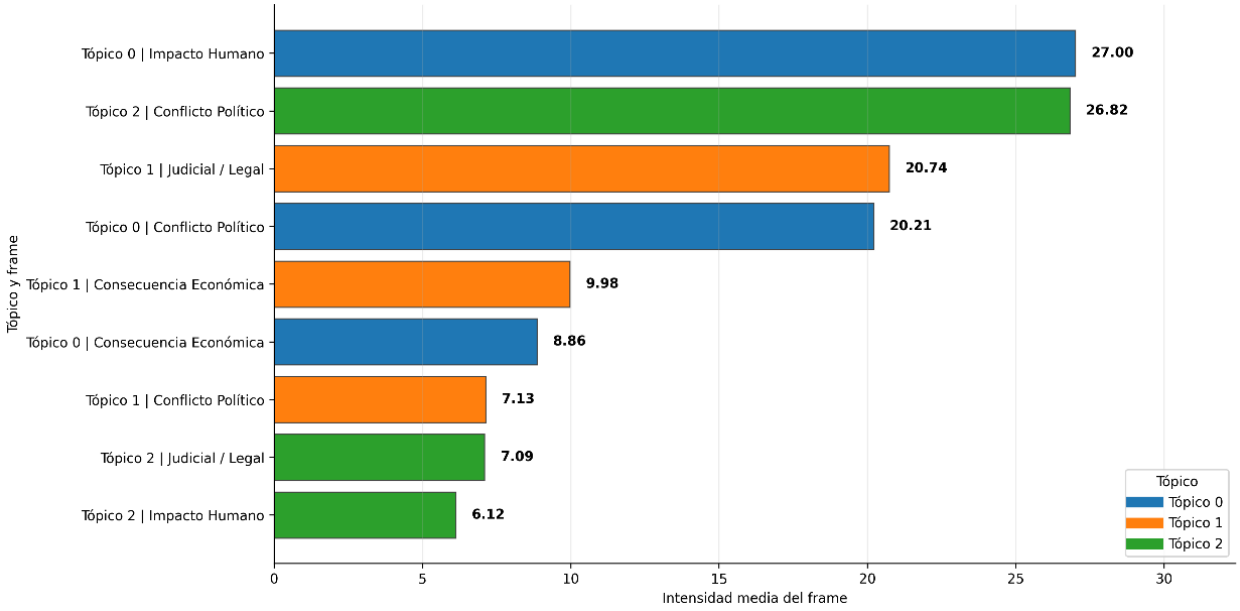
Nota. El mapa de calor muestra la intensidad relativa de cada marco narrativo en los distintos tópicos del modelo LDA.

En conjunto, estos resultados indican que los marcos narrativos constituyen una dimensión interpretativa adicional al contenido temático, permitiendo distinguir no solo qué se discute, sino cómo se presenta el fenómeno.

Frente a los marcos narrativos por tópico, se identifica que el Tópico 0 se asocia principalmente con impacto humano y conflicto político, el Tópico 1 con el marco judicial/legal, y el Tópico 2 con conflicto político.

Esta concentración sugiere la recurrencia de encuadres interpretativos específicos según el tipo de contenido abordado.

Figura 39. Marcos narrativos dominantes por tópico



Nota. La figura resume los marcos con mayor peso relativo en cada eje temático.

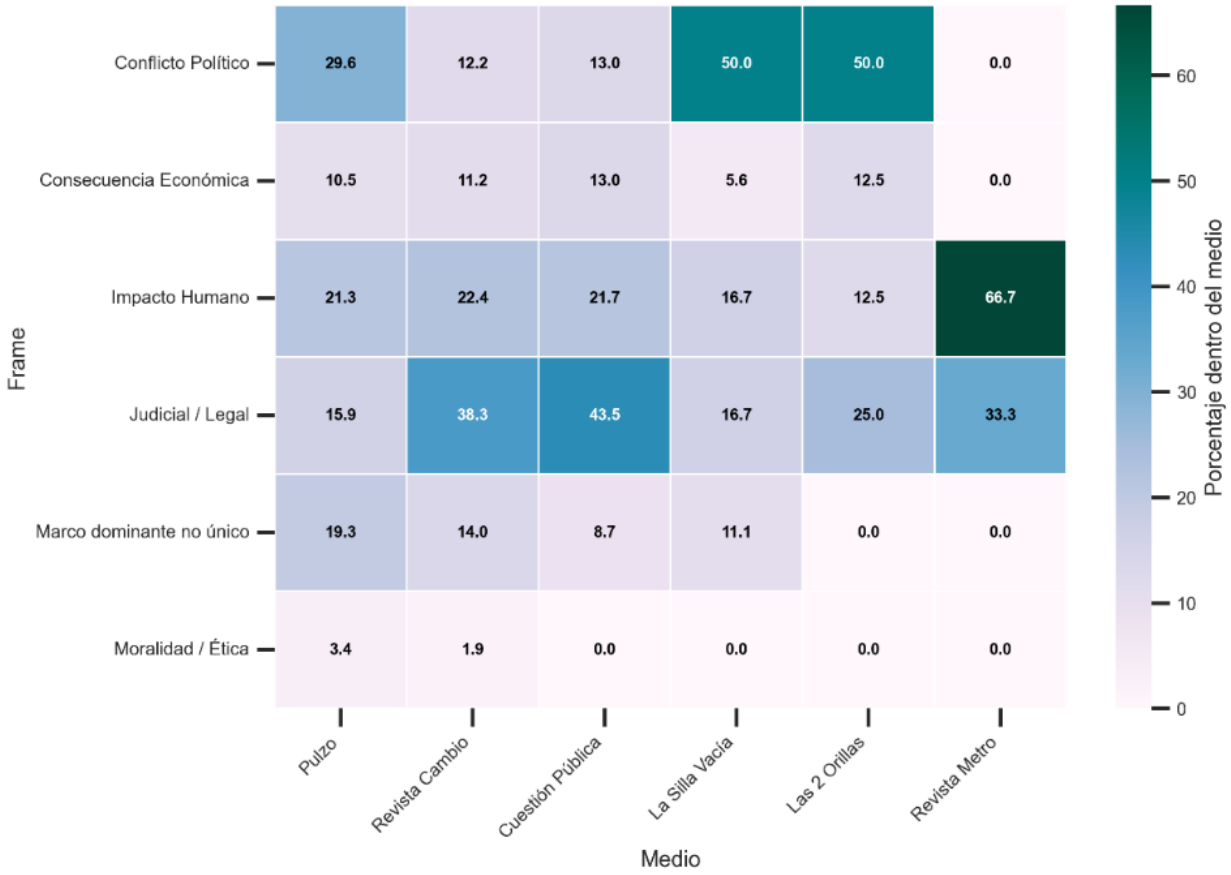
8.8.2. Distribución de marcos por medio

La distribución de marcos por medio revela variaciones sustanciales en los encuadres predominantes. Algunos medios presentan mayor énfasis en el marco judicial/legal particularmente Revista Cambio y Cuestión Pública, mientras que otros destacan por la centralidad del conflicto político, como Pulzo, La Silla Vacía y Las 2 Orillas.

El marco de impacto humano alcanza valores especialmente elevados en Revista Metro (66.7), lo que indica un enfoque centrado en las consecuencias sociales del fenómeno. En contraste, el marco de moralidad/ética presenta niveles bajos en todos los medios, sugiriendo que la cobertura privilegia dimensiones políticas, institucionales o legales por encima de evaluaciones normativas explícitas.

Estas diferencias reflejan estrategias narrativas diferenciadas dentro del ecosistema mediático analizado.

Figura 40. Distribución de marcos narrativos según medio digital

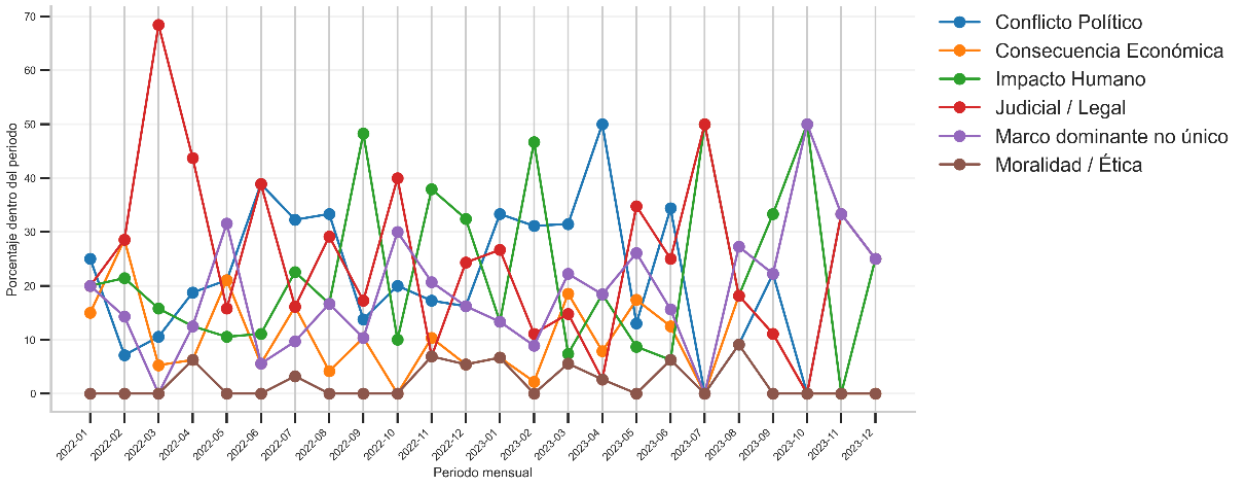


Nota. El mapa de calor muestra la intensidad relativa de cada marco narrativo en los distintos medios analizados.

8.8.3. Evolución temporal de los marcos

La evolución temporal muestra variaciones significativas en el comportamiento de los marcos a lo largo del periodo analizado. El marco judicial/legal presenta picos pronunciados en varios momentos, alcanzando valores cercanos al 70 %, lo que indica episodios de cobertura centrados en procesos judiciales.

Figura 41. Evolución temporal de los marcos narrativos



Nota. La figura muestra la variación temporal del peso relativo de los distintos marcos a lo largo del periodo analizado.

El conflicto político también exhibe fluctuaciones relevantes con incrementos abruptos en determinados meses, mientras que el impacto humano mantiene una presencia relativamente constante con oscilaciones moderadas. El marco de moralidad/ética permanece marginal durante todo el periodo.

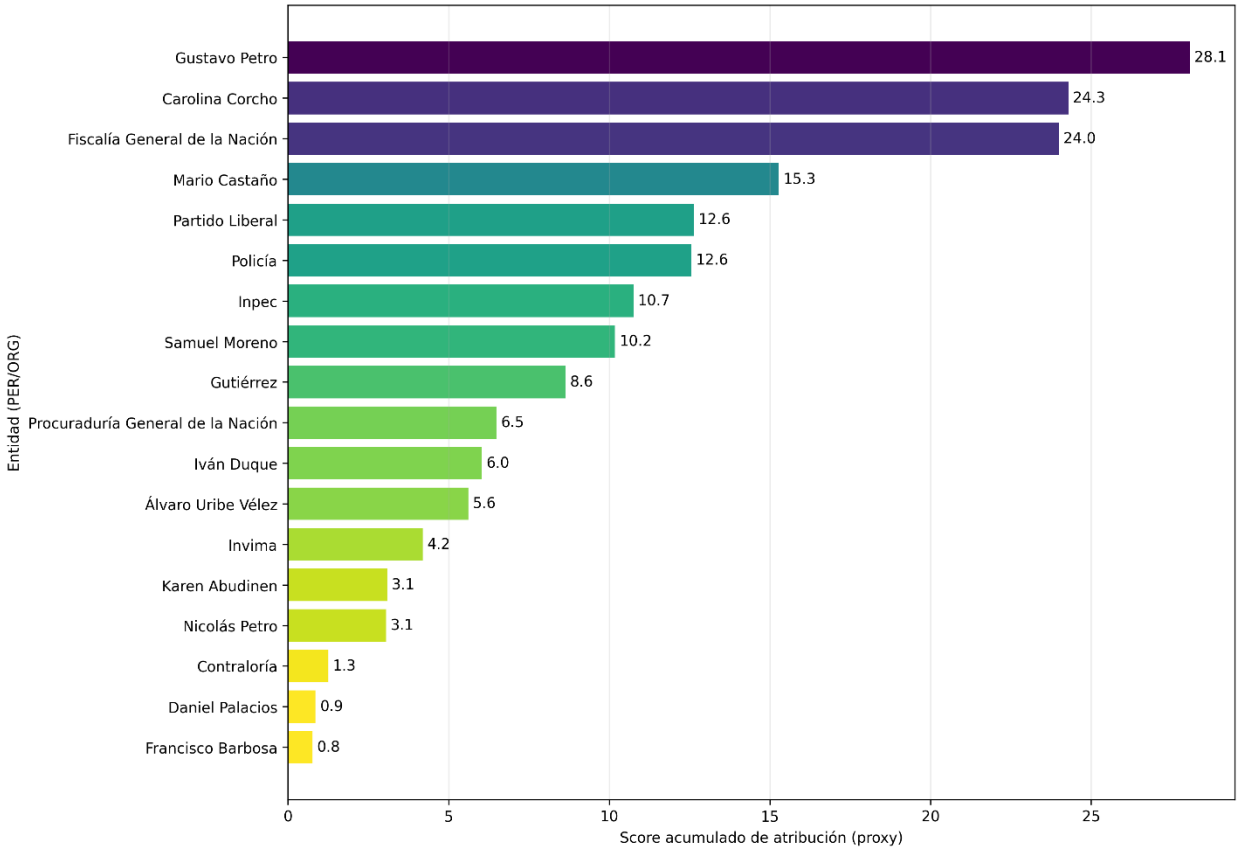
La coexistencia de picos puntuales y tendencias persistentes sugiere que la cobertura responde tanto a eventos coyunturales como a patrones narrativos relativamente estables.

8.8.4. Atribución discursiva de responsabilidad

La atribución de responsabilidad se concentra en un conjunto reducido de actores políticos e institucionales. El mayor puntaje corresponde a Gustavo Petro (28.1), seguido por Carolina Corcho (24.3) y la Fiscalía General de la Nación (24), lo que indica una focalización significativa en figuras gubernamentales y organismos estatales.

También aparecen instituciones como la Policía, el Inpec y partidos políticos, lo que sugiere una combinación de responsabilización individual e institucional.

Figura 42. Entidades con mayor atribución de responsabilidad

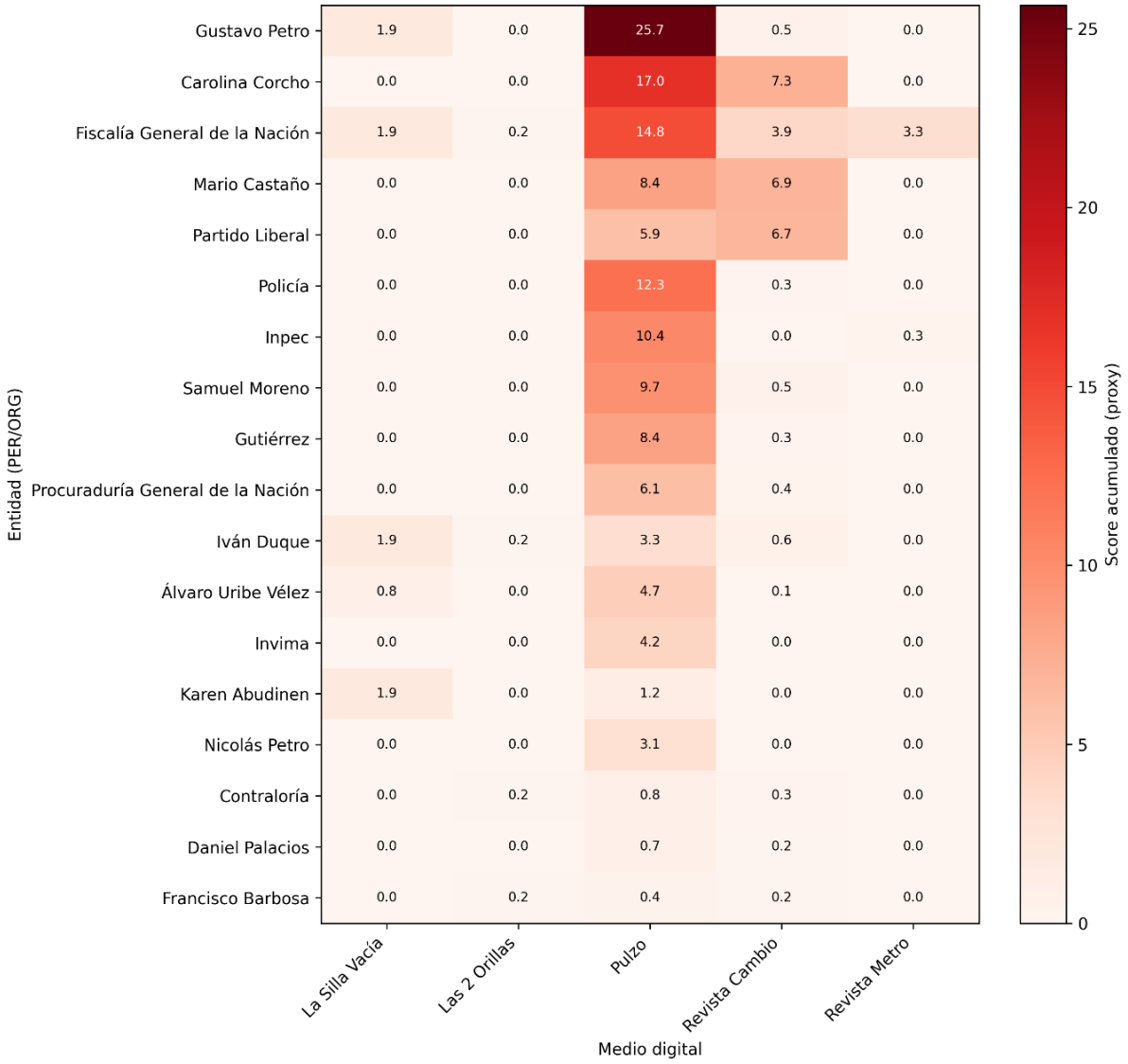


Nota. La figura muestra las entidades más frecuentemente asociadas a responsabilidad en los textos analizados.

La distribución por medio evidencia patrones diferenciados de focalización únicamente entre los medios que exhiben atribución explícita de responsabilidad según los criterios analíticos definidos. En este estudio, dicha atribución se operacionaliza mediante la identificación de documentos con alta negatividad emocional, presencia de entidades reconocidas y uso de léxico acusatorio asociado a imputación directa. Algunos medios concentran la atribución en actores gubernamentales específicos, mientras que otros la distribuyen entre múltiples instituciones.

La ausencia de determinados medios en este análisis no implica falta de cobertura del tema, sino ausencia de patrones de responsabilización explícita detectables mediante el enfoque léxico-computacional empleado.

Figura 43. Distribución de responsabilidad por medio

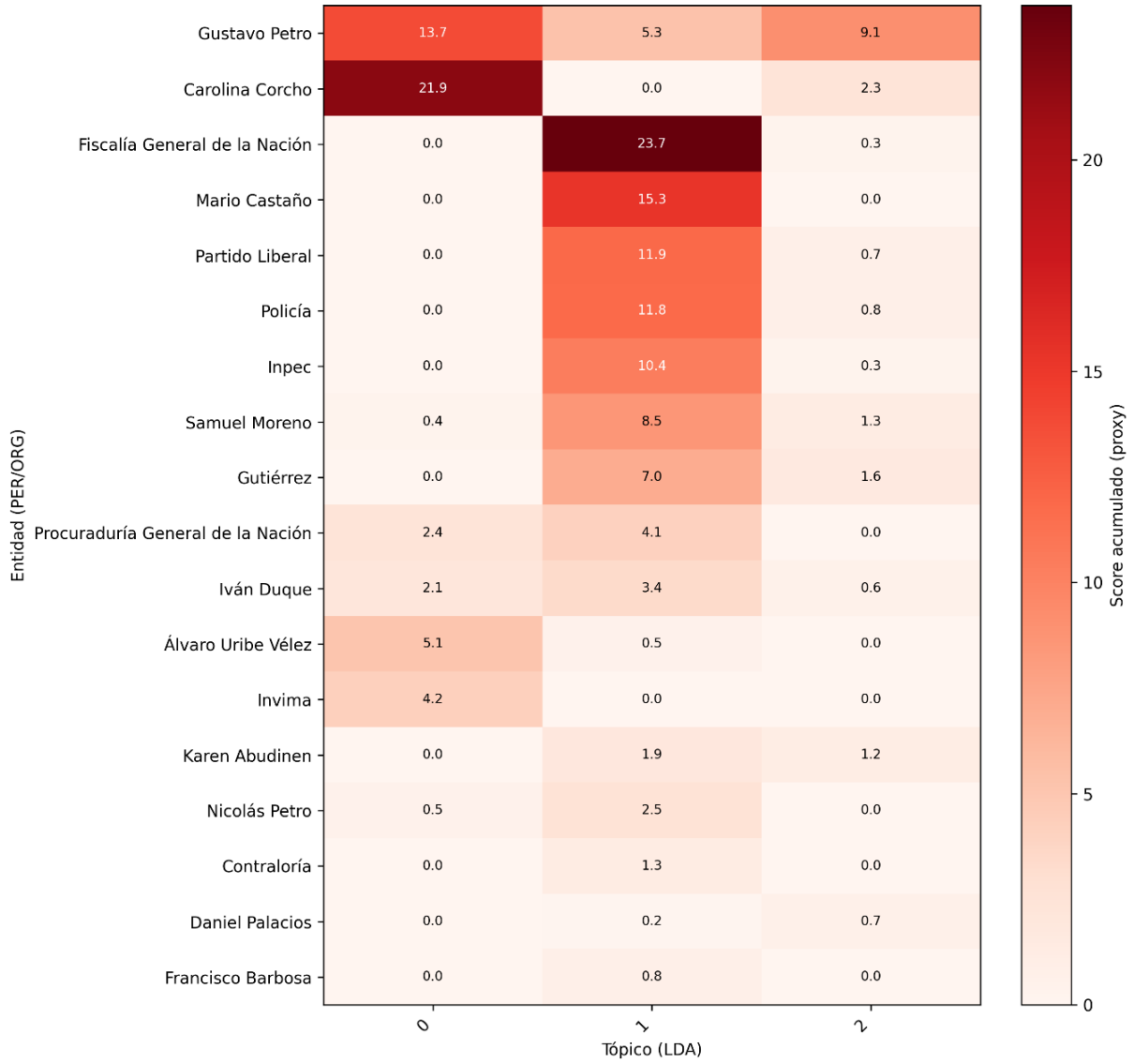


Nota. El mapa de calor muestra la frecuencia con la que cada medio asocia responsabilidad a distintas entidades.

La relación entre entidades y tópicos indica que la atribución depende del eje temático dominante. El Tópico 1 concentra la mayor intensidad de responsabilización para la mayoría de los actores, mientras que los Tópicos 0 y 2 presentan distribuciones más dispersas.

Este patrón sugiere que la personalización del discurso varía según el tipo de narrativa predominante en cada tema, especialmente en aquellos asociados a procesos institucionales o políticos.

Figura 44. Distribución de responsabilidad por eje temático



Nota. La figura muestra la relación entre entidades atribuidas como responsables y los tópicos del modelo LDA.

Por tanto, los resultados de este componente no deben interpretarse como imputación factual o judicial de culpa, sino como una aproximación computacional a patrones de responsabilización discursiva en la cobertura mediática.

8.9. Síntesis integradora de resultados

El análisis computacional del discurso mediático sobre corrupción en el sector salud en Colombia (2022–2023) permitió identificar patrones estructurales consistentes en las dimensiones temática, emocional y narrativa del corpus analizado. La convergencia de resultados provenientes del modelado temático, el análisis de sentimiento, la detección de marcos narrativos y la atribución de responsabilidad indica la existencia de una organización discursiva sistemática y no aleatoria.

En la dimensión temática, el modelo LDA optimizado identificó tres ejes semánticos coherentes que estructuran el corpus. La estabilidad de esta configuración fue corroborada mediante métricas de coherencia, análisis de generalización y triangulación con modelos alternativos no supervisados (HDP y BERTopic), lo que respalda que la estructura obtenida refleja regularidades semánticas del discurso mediático y no artefactos metodológicos.

En la dimensión emocional, el análisis de sentimiento evidencia un predominio sostenido de polaridad negativa a lo largo del periodo estudiado. La utilización conjunta de un modelo supervisado basado en Transformers (RoBERTa) y recursos léxicos especializados permitió verificar la consistencia de esta tendencia y examinar su variabilidad entre medios y tópicos.

El análisis de marcos narrativos muestra que la cobertura se articula mediante encuadres interpretativos recurrentes que enfatizan dimensiones específicas del fenómeno, como responsabilidad institucional, conflicto político, judicialización o impacto social. Estos marcos presentan asociaciones diferenciadas tanto con los ejes temáticos como con los medios, lo que evidencia patrones editoriales distintos en la construcción discursiva del problema.

La atribución discursiva de responsabilidad se concentra en un conjunto limitado de actores institucionales y políticos, lo que sugiere una focalización del discurso en responsables

identificables. Esta concentración varía según el medio y el eje temático, indicando que la asignación de responsabilidad forma parte de las estrategias narrativas mediante las cuales se interpreta el fenómeno.

En conjunto, los hallazgos evidencian que la cobertura mediática del periodo analizado se caracteriza por una agenda temática relativamente concentrada, un tono emocional predominantemente negativo y la presencia de marcos narrativos estables que orientan la interpretación de los hechos. La coherencia entre métodos analíticos independientes aporta evidencia sólida de la consistencia de estos patrones a gran escala.

Desde el punto de vista metodológico, el pipeline reproducible de Procesamiento de Lenguaje Natural demostró su capacidad para integrar técnicas de modelado temático, análisis de sentimiento y análisis narrativo dentro de un marco analítico coherente y trazable, en concordancia con el enfoque CRISP-DM aplicado al estudio.

En síntesis, los resultados apoyan la hipótesis alternativa (H_1), al evidenciar la existencia de estructuras temáticas coherentes y variaciones relevantes en el tono emocional del discurso entre los medios digitales analizados.

8.10. Propuesta de solución a la problemática

Los resultados obtenidos evidencian que el discurso mediático sobre corrupción en el sector salud configura interpretaciones sociales del fenómeno mediante patrones temáticos, emocionales y narrativos recurrentes. Esta constatación permite trascender el análisis descriptivo y plantea la posibilidad de desarrollar herramientas aplicadas orientadas a mejorar la transparencia informativa y el control social.

La investigación no se limita a caracterizar el discurso, sino que aporta evidencia empírica para diseñar mecanismos que faciliten la comprensión pública del fenómeno y reduzcan las asimetrías informacionales entre instituciones, medios y ciudadanía.

Asimismo, la complejidad del problema exige instrumentos capaces de integrar múltiples dimensiones del discurso contenido temático, tono emocional, marcos interpretativos y actores involucrados dentro de un sistema analítico coherente y sostenible en el tiempo.

Con base en estos elementos, se presenta una propuesta estructurada en tres componentes: situación actual, oportunidades derivadas de los hallazgos y solución orientada a fortalecer la transparencia informativa y el control social.

8.10.1. Situación actual

La información mediática sobre corrupción en salud se caracteriza por su fragmentación, diversidad de enfoques editoriales y predominio de narrativas centradas en actores específicos o eventos coyunturales. Esta configuración dificulta la comprensión integral del fenómeno y limita la capacidad de la ciudadanía para ejercer un control informado.

Aunque los medios cumplen un papel esencial en la visibilización de irregularidades, la cobertura no necesariamente ofrece una visión sistemática de las causas estructurales del problema ni de su impacto global. La heterogeneidad discursiva puede generar interpretaciones divergentes entre audiencias y dificultar la construcción de diagnósticos compartidos.

Adicionalmente, no existen mecanismos ampliamente disponibles que permitan monitorear de forma continua cómo evoluciona el discurso público sobre corrupción en salud, lo que restringe la capacidad de organizaciones sociales e instituciones para anticipar cambios en la agenda informativa o evaluar tendencias comunicativas.

8.10.2. Oportunidades

Las técnicas de Procesamiento de Lenguaje Natural permiten analizar grandes volúmenes de contenido mediático y extraer patrones estructurales del discurso con alto grado de reproducibilidad.

El modelado temático facilita la identificación de agendas informativas dominantes y su evolución temporal; el análisis de sentimiento permite estimar la intensidad emocional del

debate; la detección de marcos narrativos revela los encuadres interpretativos predominantes; y el análisis de actores permite mapear la visibilidad y atribución discursiva de responsabilidad.

Estas capacidades ofrecen una base tecnológica para transformar información dispersa en conocimiento útil para la ciudadanía, contribuyendo al fortalecimiento de la transparencia y la vigilancia democrática.

8.10.3. Propuesta de solución al problema planteado

En el ordenamiento jurídico colombiano, la participación ciudadana en la vigilancia de la gestión pública cuenta con un sólido fundamento constitucional y legal, materializado en mecanismos como las veedurías ciudadanas, concebidas para ejercer control social sobre la administración del Estado y la ejecución de recursos públicos. Este marco reconoce que el acceso oportuno a información clara y verificable constituye una condición esencial para el ejercicio efectivo de la transparencia y la rendición de cuentas. Sin embargo, en contextos caracterizados por alta complejidad informativa y volumen creciente de contenidos mediáticos digitales, la disponibilidad formal de información no siempre se traduce en capacidad real de análisis por parte de la ciudadanía.

En este escenario, la incorporación de herramientas tecnológicas avanzadas orientadas al procesamiento automatizado de grandes volúmenes de texto emerge como una oportunidad para fortalecer estos mecanismos de control social, ampliando las capacidades de monitoreo, interpretación y seguimiento del discurso público. Bajo esta perspectiva, la creación de un Observatorio Digital de Discurso Mediático sobre Corrupción se plantea como un instrumento complementario a las veedurías ciudadanas, diseñado para facilitar el acceso abierto a información estructurada y analíticamente procesada, en consonancia con los principios constitucionales de participación, transparencia y control democrático.

Esta propuesta se fundamenta en los hallazgos del estudio, que evidencian la existencia de patrones discursivos estructurados, persistencia de encuadres interpretativos específicos y

variaciones significativas entre medios en la construcción narrativa de la corrupción en el sector salud. Tales características sugieren la necesidad de mecanismos permanentes que permitan observar, comparar y contextualizar la cobertura informativa más allá de episodios coyunturales.

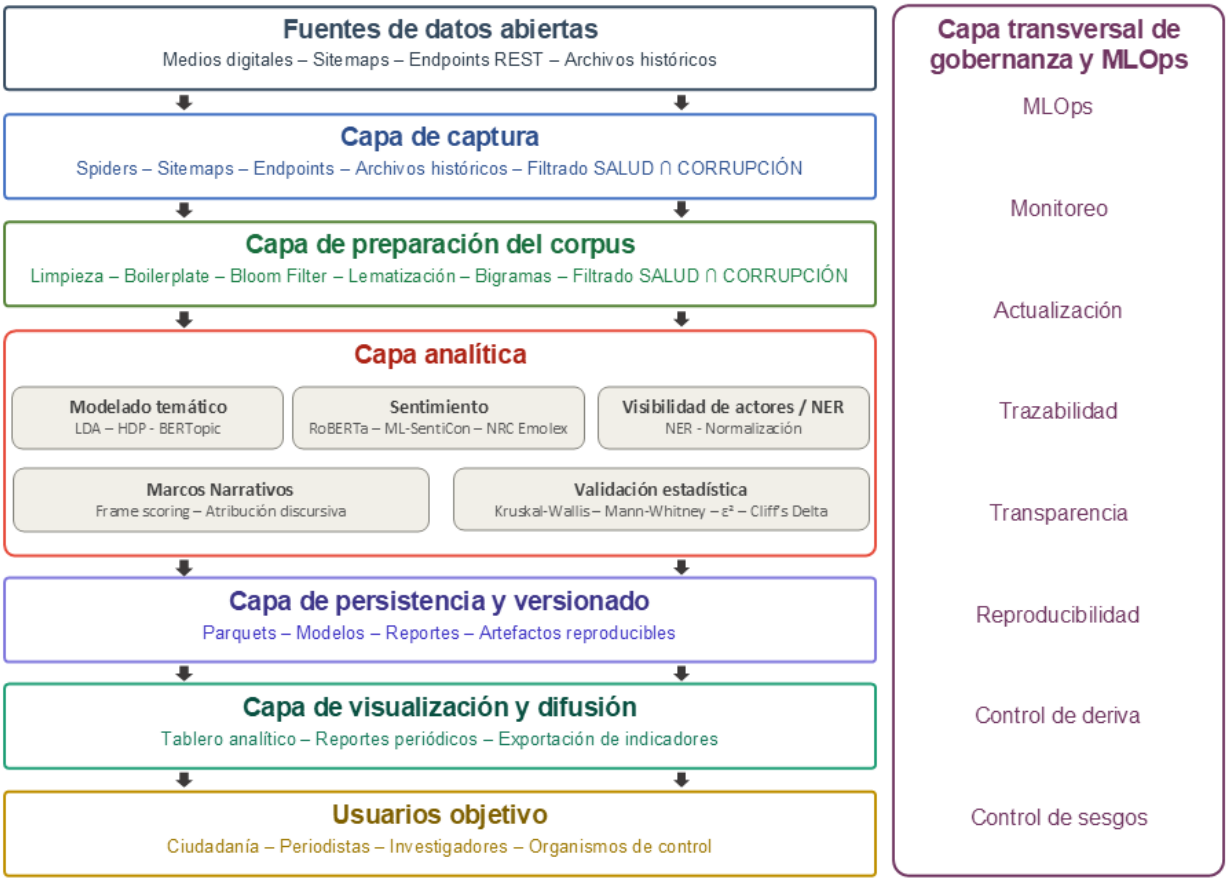
El observatorio operaría como un sistema de vigilancia discursiva basado en técnicas de Procesamiento de Lenguaje Natural, capaz de analizar de manera continua contenidos provenientes de medios digitales abiertos. Su finalidad no sería evaluar la veracidad de la información ni intervenir en el contenido editorial, sino proporcionar indicadores estructurados sobre cómo se construye mediáticamente el fenómeno.

Entre sus funciones principales se incluirían:

- Monitoreo continuo de la agenda temática mediática, identificando cambios en la prominencia de los ejes narrativos.
- Estimación periódica del tono emocional del discurso, permitiendo detectar intensificaciones o atenuaciones del tratamiento informativo.
- Identificación de marcos narrativos predominantes y su evolución temporal.
- Mapeo de actores relevantes y patrones de atribución de responsabilidad.
- Generación de visualizaciones interactivas orientadas a públicos no especializados.
- Elaboración de reportes analíticos que faciliten la interpretación contextual de los datos.

Desde el punto de vista aplicado, la viabilidad del observatorio no parte de una formulación abstracta, sino de la base técnica ya desarrollada en esta investigación mediante un pipeline computacional reproducible para recolección, depuración, modelado y análisis del corpus mediático. En consecuencia, la propuesta puede entenderse como una proyección funcional del sistema ya implementado, orientada a su operación continua sobre medios digitales abiertos o semiautomatizables, con énfasis en actualización periódica, trazabilidad analítica y disponibilidad de indicadores para distintos tipos de usuario.

Figura 45. Arquitectura técnico-operativa propuesta del Observatorio Digital de Discurso Mediático



Nota. La figura sintetiza la arquitectura técnico-operativa propuesta para el observatorio, basada en el pipeline funcional desarrollado en la investigación. La estructura integra captura automatizada, preparación del corpus, analítica discursiva, persistencia de resultados, visualización y monitoreo operativo bajo principios de trazabilidad, actualización periódica y transparencia metodológica.

Tabla 21. Viabilidad operativa y métricas de evaluación propuestas para el observatorio

Componente	Requerimiento operativo mínimo	Evidencia de viabilidad en el proyecto	Métricas sugeridas de evaluación futura
Captura de datos	Ejecución programada sobre medios abiertos o semiautomatizables, conectividad estable y control de errores de extracción	El proyecto ya implementó spiders, rutas históricas, endpoints JSON, sitemaps y estrategias de fallback según la arquitectura de cada fuente	Tasa de extracción exitosa, porcentaje de fuentes activas, latencia de actualización, porcentaje de errores por fuente
Preparación del corpus	Procesamiento automatizado de limpieza, deduplicación,	El pipeline desarrollado ya ejecuta normalización, limpieza	Completitud del corpus, tasa de documentos válidos tras limpieza,

	boilerplate y filtrado temático	estructural, deduplicación y filtrado SALUD ∩ CORRUPCIÓN	proporción de duplicados removidos, cobertura temática efectiva
Analítica temática	Entorno reproducible para reentrenamiento y validación periódica de modelos	El proyecto ya implementó LDA como modelo principal y triangulación con HDP y BERTopic	Coherencia C_v, C_npmi, Topic Diversity, estabilidad entre ejecuciones, drift temático
Analítica emocional	Ejecución periódica del modelo contextual y contraste con recursos léxicos	El proyecto ya implementó RoBERTa, ML-SentiCon y NRC EmoLex con validación cruzada	Entropía media, margen de confianza, correlación entre métodos, desacuerdo intermodelo
Actores, frames y responsabilidad	Pipeline integrado de NER, scoring de frames y atribución discursiva	El sistema ya genera visibilidad de actores, marcos narrativos y patrones de responsabilización discursiva	Cobertura de entidades, estabilidad de rankings, densidad de frames, consistencia de atribución
Persistencia y difusión	Almacenamiento versionado de artefactos y generación de salidas interpretables	El proyecto ya produce parquets, reportes, figuras y artefactos de modelos reproducibles	Disponibilidad de reportes, frecuencia de actualización, integridad de artefactos, trazabilidad de versiones
Gobernanza y sostenibilidad	Monitoreo operativo, actualización periódica y control metodológico	La propuesta ya contempla prácticas de MLOps, control de sesgos y reproducibilidad del pipeline	Tiempo de actualización, tasa de fallos de proceso, monitoreo de deriva, trazabilidad metodológica

Nota. La tabla resume condiciones mínimas de operación del observatorio a partir de los componentes ya implementados en el proyecto y propone métricas orientativas para su evaluación futura. Estas métricas no corresponden a una implementación en producción, sino a criterios de seguimiento técnico y analítico derivados de la arquitectura funcional planteada. Elaboración propia.

El sistema estaría diseñado como una plataforma abierta, accesible a ciudadanía, periodistas, investigadores y organismos de control, con el propósito de equilibrar el acceso de productores y consumidores de información. Al ofrecer evidencia empírica sobre patrones discursivos, el observatorio contribuiría a una comprensión más informada del fenómeno y a la evaluación crítica de la cobertura mediática.

Desde el punto de vista técnico, la sostenibilidad del observatorio dependería de la implementación de prácticas de MLOps que garanticen:

- Actualización periódica de los modelos analíticos frente a cambios lingüísticos y temáticos.
- Incorporación gradual de nuevas fuentes de información.
- Monitoreo continuo del desempeño de los modelos.
- Control de sesgos y degradación de resultados.

- Reproducibilidad del pipeline analítico.

Asimismo, la arquitectura del sistema debería contemplar mecanismos de gobernanza que aseguren su independencia operativa y transparencia metodológica, evitando que la herramienta sea utilizada con fines de vigilancia política o control editorial. En este sentido, el observatorio se concibe como un instrumento de apoyo a la deliberación pública y no como un mecanismo de supervisión institucional.

La implementación de esta propuesta, en articulación con instancias académicas e institucionales, permitiría transformar los hallazgos de la investigación en una herramienta práctica con potencial de impacto social, al facilitar el seguimiento transversal del discurso mediático sobre corrupción y promover una mayor responsabilidad informativa. En términos aplicados, el observatorio contribuiría a fortalecer la capacidad de la sociedad para interpretar críticamente la información disponible, favoreciendo procesos de transparencia, participación ciudadana y control social fundamentados en evidencia empírica. Asimismo, al ofrecer resultados mediante interfaces accesibles y comprensibles, reduciría las barreras técnicas asociadas al uso de herramientas avanzadas de análisis de datos, ampliando la disponibilidad y el aprovechamiento de información estructurada por parte de ciudadanía, periodistas, investigadores y tomadores de decisión.

En este sentido, el observatorio no se plantea como una idea desvinculada del desarrollo empírico de la investigación, sino como una extensión aplicada del sistema analítico ya construido. Su valor agregado radica en transformar un pipeline reproducible de investigación en una infraestructura de monitoreo discursivo continuo, con potencial para fortalecer el control social, la deliberación pública informada y la transparencia en el seguimiento mediático de la corrupción en salud. Así, el aporte aplicado del estudio no se limita a la formulación conceptual del observatorio, sino que se sustenta en la demostración previa de viabilidad técnica, trazabilidad metodológica y generación efectiva de indicadores analíticos verificables sobre el corpus examinado.

9. Discusión

Esta sección integra los hallazgos empíricos obtenidos mediante el modelado temático LDA optimizado (k = 3), el análisis de sentimiento basado en modelos Transformer (RoBERTa) y la triangulación metodológica con modelos alternativos, en articulación con los enfoques teóricos del framing mediático y la agenda-setting desarrollados en el marco conceptual. En conjunto, los resultados indican que la cobertura digital sobre la corrupción en el sector salud en Colombia presenta patrones discursivos estructurados, diferenciados entre medios y modulados emocionalmente.

Desde una perspectiva interpretativa, la existencia de patrones temáticos y emocionales diferenciados entre medios sugiere que la representación mediática del fenómeno responde a procesos de selección y énfasis propios del sistema informativo. Este comportamiento es consistente con la literatura sobre agenda-setting, según la cual los medios contribuyen a la construcción de interpretaciones públicas sobre determinados problemas y perspectivas.

Los resultados evidencian asimetrías de framing y polarización negativa selectiva que no se distribuyen de manera homogénea ni entre los ejes narrativos ni entre los medios analizados. Esta variabilidad indica estilos editoriales diferenciados en la construcción del problema, coherentes con la noción de framing selectivo planteada por la literatura especializada.

En particular, los ejes narrativos asociados a la confrontación política, la judicialización y la atribución de responsabilidades institucionales presentan mayores niveles de polaridad negativa, mientras que el eje vinculado a la crisis estructural del sistema de salud tiende a adoptar registros comparativamente más descriptivos. Este contraste sugiere que la corrupción en salud es frecuentemente interpretada dentro de narrativas de conflicto político más amplias, desplazando parcialmente el foco desde dimensiones técnico-administrativas hacia disputas institucionales.

Asimismo, el modelado temático evidencia una diferenciación semántica entre el eje narrativo de la crisis estructural del sistema de salud y aquellos centrados en la confrontación política y la justicia. El análisis longitudinal sugiere que, en determinados periodos, el léxico técnico-sanitario aparece integrado en narrativas de carácter político. Este patrón puede interpretarse como una reconfiguración del marco interpretativo, en la cual los problemas estructurales del sistema se discuten en conexión con disputas institucionales más amplias. Desde la teoría del framing, ello implica la articulación de distintos dominios discursivos dentro de una misma narrativa mediática.

Otro hallazgo relevante es la recurrencia de referencias a casos emblemáticos anteriores al periodo analizado. Esta dinámica sugiere la existencia de continuidad narrativa, mediante la cual episodios previos son utilizados como contexto interpretativo para nuevas denuncias. Aunque esta estrategia puede reforzar la percepción de recurrencia del problema, también puede contribuir a la homogenización discursiva de fenómenos heterogéneos, dado que la corrupción en el sector salud presenta manifestaciones institucionales y territoriales diversas.

Además de los encuadres políticos y judiciales, se identifica un eje narrativo centrado en las consecuencias sociales de la corrupción, particularmente sobre poblaciones vulnerables. Este hallazgo evidencia la presencia de un marco orientado a los impactos sociales del fenómeno, que desplaza el foco desde los actores responsables hacia sus efectos en la ciudadanía. Aunque este eje presenta menor peso relativo frente a las narrativas políticas dominantes, su presencia sostenida indica que la cobertura mediática incorpora una dimensión social y moral en la interpretación del problema.

En conjunto, la evidencia sugiere que el discurso mediático sobre corrupción en el sector salud en Colombia se configura como un campo narrativo heterogéneo, donde coexisten marcos políticos, judiciales, estructurales y sociales con comportamientos diferenciados según el medio y el contexto temporal. La integración de modelado temático, análisis de sentimiento y

análisis narrativo permite comprender no solo qué temas se abordan, sino también cómo se interpretan discursivamente.

Estos resultados son consistentes con la idea de que los medios desempeñan un papel relevante en la configuración del debate público sobre la corrupción en el sector salud, contribuyendo a la construcción de significados colectivos, a la visibilización de determinados actores y a la priorización de ciertos aspectos dentro de la agenda informativa, sin que ello implique inferir efectos directos sobre las percepciones individuales de la ciudadanía.

10. Conclusiones y Trabajo Futuro

10.1. Conclusiones

La presente investigación evidencia que el discurso mediático digital sobre la corrupción en el sector salud en Colombia durante el periodo 2022–2023 presenta una organización estructurada identificable mediante técnicas de Procesamiento de Lenguaje Natural aplicadas a gran escala. Los resultados descartan una distribución aleatoria del contenido y confirman la existencia de patrones temáticos, lingüísticos y emocionales recurrentes en la cobertura informativa.

Desde la dimensión temática, el modelado LDA optimizado permitió identificar tres ejes narrativos diferenciados que estructuran el corpus analizado. Estos ejes representan dimensiones institucionales, políticas y sociales del fenómeno, lo que indica que la corrupción en salud es abordada mediáticamente mediante múltiples encuadres interpretativos. Este resultado es consistente con los postulados del framing mediático y confirma la complejidad semántica del problema.

En la dimensión emocional, el análisis de sentimiento evidencia un predominio sostenido de polaridad negativa, con variaciones consistentes entre medios y ejes temáticos. La convergencia entre un modelo contextual basado en Transformers y recursos léxicos especializados respalda la consistencia de este patrón. Asimismo, el análisis de atribución muestra una concentración discursiva de la responsabilidad en actores institucionales y políticos específicos.

Desde el punto de vista metodológico, el pipeline desarrollado demuestra que la integración de modelado temático probabilístico, análisis de sentimiento basado en embeddings y evaluación cuantitativa mediante métricas de coherencia, generalización y comparación entre modelos permite examinar fenómenos sociopolíticos complejos de manera reproducible y

escalable. Este enfoque constituye un aporte metodológico transferible a otros estudios de análisis del discurso mediático en grandes volúmenes de texto.

En conjunto, los hallazgos indican que la cobertura mediática del periodo analizado se caracteriza por una agenda temática concentrada, un tono emocional predominantemente negativo y marcos narrativos recurrentes que orientan la interpretación del fenómeno. Estos resultados apoyan la hipótesis alternativa (H_1) al evidenciar la existencia de estructuras discursivas coherentes y variaciones relevantes entre medios digitales.

Aportes teóricos, metodológicos y aplicados

En el plano teórico, la investigación aporta evidencia empírica que respalda las teorías de framing y agenda-setting mediante su operacionalización computacional a gran escala.

En el plano metodológico, propone un flujo de trabajo reproducible que integra técnicas de PLN probabilístico y modelos basados en Transformers, adaptable a otros dominios temáticos y contextos geográficos.

Desde una perspectiva aplicada, los resultados ofrecen insumos para el análisis crítico de la cobertura mediática y para procesos de monitoreo informativo en sectores estratégicos como la salud, así como para entornos institucionales orientados al análisis de grandes volúmenes de información textual.

Limitaciones de la investigación

Se reconocen limitaciones asociadas tanto al diseño como a las fuentes utilizadas. La exclusión de contenidos protegidos por suscripción puede generar sesgos de representatividad hacia medios de acceso abierto. Asimismo, las restricciones de longitud de los modelos Transformer implicaron priorizar segmentos textuales más informativos, lo que podría omitir matices presentes en artículos extensos.

El carácter dinámico del lenguaje mediático digital exige procesos continuos de limpieza y normalización, lo que dificulta la construcción de modelos completamente generalizables. Adicionalmente, las diferencias entre métodos contextuales y léxicos reflejan la complejidad del

discurso político, donde la valoración negativa puede expresarse de forma indirecta. La triangulación metodológica permitió mitigar estos sesgos y fortalecer la validez de los resultados.

Asimismo, se reconoce como limitación que el etiquetado conceptual de los tópicos fue realizado por un único investigador, sin validación inter-jueces ni revisión experta formal. Aunque esta fase se apoyó en criterios explícitos y en controles metodológicos indirectos, una validación interpretativa con codificadores independientes fortalecería la robustez analítica de futuras aplicaciones.

Cierre conclusivo

En síntesis, el estudio confirma que el análisis computacional del discurso mediático constituye una herramienta robusta para examinar la construcción discursiva de fenómenos complejos como la corrupción en el sector salud. La integración de rigor metodológico, validación empírica y marco teórico permite avanzar hacia enfoques interdisciplinarios que articulan ciencia de datos y ciencias sociales.

10.2. Trabajo futuro

Futuras investigaciones podrían ampliar el alcance temporal del análisis para evaluar la evolución del discurso en periodos más extensos o en contextos comparativos. Asimismo, sería pertinente incorporar otras fuentes informativas, como redes sociales o documentos institucionales, para obtener una visión más completa del ecosistema comunicativo.

También se podrían explorar modelos semánticos más avanzados y técnicas de análisis multimodal que integren texto, imágenes y otros formatos de contenido digital.

Finalmente, la implementación práctica del observatorio propuesto permitiría evaluar su utilidad en escenarios reales de transparencia y control social, así como ajustar sus componentes técnicos mediante procesos iterativos de mejora continua.

10.3. Declaración de uso de herramientas de inteligencia artificial

Durante la elaboración del presente trabajo se utilizaron herramientas de inteligencia artificial generativa (incluyendo ChatGPT, Gemini, Claude y Perplexity) exclusivamente como apoyo para la revisión de redacción, mejora del estilo académico, organización del contenido y orientación técnica puntual.

La formulación del problema, el diseño metodológico, la recolección y procesamiento de datos, los análisis realizados, la interpretación de resultados y las conclusiones son de autoría propia del investigador. Las herramientas de IA no fueron empleadas para generar datos, ejecutar el pipeline analítico ni producir resultados empíricos.

El uso de estas tecnologías se enmarca en prácticas de apoyo instrumental, sin delegación de la responsabilidad intelectual ni académica del trabajo presentado.

11. Referencias

Adam, I., & Fazekas, M. (2021). Are emerging technologies helping win the fight against corruption? A review of the state of evidence. *Information Economics and Policy*, *57*, 100950. https://doi.org/10.1016/j.infoecopol.2021.100950

Adcock, R., & Collier, D. (2016). *Measurement Validity: A Shared Standard for Qualitative and Quantitative Research*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199384426.003.0002

Arroyave, J., & Barrios, M. (2023). Narrativas mediáticas sobre la corrupción política en Colombia: análisis de prensa digital 2018–2022. *Palabra Clave*, *26*(2), 1-24. https://doi.org/10.5294/pacla.2023.26.2.4

Asociación Colombiana de Medios de Información. (2023). *Portal oficial*. https://www.ami.org.co/

Association of Internet Researchers. (2019). *Internet Research: Ethical Guidelines 3.0*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 610-623. https://doi.org/10.1145/3442188.3445922

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993-1022. https://jmlr.org/papers/v3/blei03a.html

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 5454-5476. https://doi.org/10.18653/v1/2020.acl-main.485

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of the Biennial GSCL Conference 2009: From Form to Meaning: Processing Texts Automatically*, 31-40.

Boumans, J. W., & Trilling, D. (2016). Taking Stock of the Toolkit: An Overview of Relevant Automated Content Analysis Approaches and Techniques for Digital Journalism Scholars. *Digital Journalism*, *4*(1), 8-23. https://doi.org/10.1080/21670811.2015.1096598

Boydstun, A. E., & Shafer, H. F. (2017). The Real-World Consequences of Framing. En K. Kenski & K. H. Jamieson (Eds.), *The Oxford Handbook of Political Communication*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.46

Camacho-Collados, J., & Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research*, *63*(1), 743-788. https://doi.org/10.1613/jair.1.11259

Castells, M. (2009). *Communication Power*. Oxford University Press.

Contraloría General de la República. (2023). *Informe especial de control fiscal al sector salud 2020-2023*. https://www.contraloria.gov.co/documents/20181/0/Informe+Especial+Sector+Salud+2020-2023.pdf

Couldry, N., & Hepp, A. (2017). *The Mediated Construction of Reality*. Polity Press. https://www.politybooks.com/bookdetail/?isbn=9780745681306

Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4.ª ed.). SAGE Publications.

de Vreese, C. H. (2019). News framing: Theory and typology. *Information, Communication & Society*, *22*(6), 915-932. https://doi.org/10.1080/1369118X.2019.1576862

Departamento Administrativo Nacional de Estadística - DANE. (2023). *Proyecciones de población 2018–2042 por área y municipio (Archivo PPED-AreaMun-2018-2042_VP.xlsx)*. https://microdatos.dane.gov.co/index.php/catalog/792

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 4171-4186. https://doi.org/10.48550/arXiv.1810.04805

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2788-2797. http://proceedings.mlr.press/v119/dieng20a.html

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT Press.

El Espectador. (2023a). *Corrupción en el sistema de salud: entre sobrecostos y opacidad institucional*. https://www.elespectador.com/salud/corrupcion-en-el-sistema-de-salud-entre-sobrecostos-y-opacidad-institucional/

El Espectador. (2023b). *Los escándalos de corrupción que marcaron el gobierno Duque y el inicio del gobierno Petro*. https://www.elespectador.com/politica/los-avances-de-duque-y-el-reto-para-petro-en-la-lucha-contra-la-corrupcion/

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51-58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Fairclough, N. (2013). *Critical Discourse Analysis: The Critical Study of Language* (2.ª ed.). Routledge. https://www.routledge.com/Critical-Discourse-Analysis-The-Critical-Study-of-Language/Fairclough/p/book/9781405858229

Fan, R., Tan, C., Lim, E.-P., & Ong, D. (2022). Modeling emotion dynamics in social media discourse: Sentiment and polarization in online news. *Information Processing & Management*, *59*(4), 102972. https://doi.org/10.1016/j.ipm.2022.102972

Gaitán, L. F., Restrepo, J., & Guzmán, C. (2020). Corrupción, confianza y legitimidad institucional en América Latina: evidencias comparadas y reflexiones para Colombia. *Revista de Economía Institucional*, *22*(43), 71-98. https://doi.org/10.18601/01245996.v22n43.05

Goffman, E. (1974). *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press.

Greussing, E., & Boomgaarden, H. G. (2016). Framing the Crisis: The Role of Media Framing in the EU Crisis and Its Effects on Public Opinion. *Journal of European Public Policy*, *24*(1), 105-126. https://doi.org/10.1080/13501763.2016.1164745

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267-297. https://doi.org/10.1093/pan/mps028

Grootendorst, M. (2022). *BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure*. https://arxiv.org/abs/2203.05794

Guerrero, R., Gallego, R., & Rodríguez, D. (2019). La corrupción en salud en Colombia: ¿Qué sabemos y qué falta por saber? *Revista Gerencia y Políticas de Salud*, *18*(36), 1-14. https://doi.org/10.11144/Javeriana.rgyps18.cscs

Hernández-Sampieri, R., Fernández-Collado, C., & Baptista-Lucio, P. (2018). *Metodología de la investigación* (6.ª ed.). McGraw-Hill.

Hubert, L., & Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, *2*(1), 193-218. https://doi.org/10.1007/BF01908075

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469-485. https://doi.org/10.1080/13645579.2018.1484990

Jain, S., Mishra, D., Gupta, R., & Alvi, M. (2022). Mining Textual Data for Public Health and Corruption Analysis: A Case Study of India. *Information Systems Frontiers*, 1-19. https://doi.org/10.1007/s10796-022-10289-4

La República. (2023). *Denuncias y procesos por corrupción en la red hospitalaria colombiana*. La República. https://www.larepublica.co/economia/denuncias-y-procesos-por-corrupcion-en-la-red-hospitalaria-colombiana-3660218

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, 530-539. https://doi.org/10.3115/v1/E14-1056

Li, Y., Adams, J., Niezen, G., Tang, J., & Johnson, H. (2020). Detection of self-reported experiences with corruption on Twitter using unsupervised machine learning. *Social Sciences & Humanities Open*, *2*(1), 100060. https://doi.org/10.1016/j.ssaho.2020.100060

Lindgren, S. (2022). Data-driven discourse analysis: Using NLP for the study of media and political communication. *Social Media + Society*, *8*(2), 1-13. https://doi.org/10.1177/20563051221089545

Liu, B. (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* (2.ª ed.). Cambridge University Press. https://doi.org/10.1017/9781108639419

López-Londoño, C., & Molinares, J. (2021). Cobertura digital de la corrupción en Colombia: encuadres, actores y discursos. *Revista de Comunicación y Ciudadanía Digital*, *8*(2), 55-79. https://doi.org/10.26441/RC8.2-2021

McCombs, M. E., & Shaw, D. L. (1972). The Agenda-Setting Function of Mass Media. *Public Opinion Quarterly*, *36*(2), 176-187. https://doi.org/10.1086/267990

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment Analysis Algorithms and Applications: A Survey. *Ain Shams Engineering Journal*, *5*(4), 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

Ministerio de Salud y Protección Social. (2023). *Proyecto de ley de reforma al sistema de salud (texto radicado)*. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/proyecto-ley-reforma-salud-msps.pdf

Ministerio de Salud y Protección Social de Colombia. (1993). *Resolución 8430 de 1993: Por la cual se establecen las normas científicas, técnicas y administrativas para la investigación*

en salud. https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-8430-de-1993.pdf

Moyano, A., & Salazar, D. (2022). Transformación digital y agenda mediática en Colombia: evolución del ecosistema informativo 2015–2022. *Revista Anagramas*, *21*(40), 45-66. https://doi.org/10.22395/angr.v21n40a3

Neil Patel Digital. (2025). *Ubersuggest – Website Traffic Checker*. https://neilpatel.com/ubersuggest/

Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, *334*(6060), 1226-1227. https://doi.org/10.1126/science.1213847

Pérez, C. (2023). Análisis crítico del discurso y framing para una propuesta metodológica. *Cuaderno 198* | *Centro de Estudios en Diseño y Comunicación*, *198*, 53-63. https://doi.org/10.18682/cdc.vi198.9819

Revista P&M. (2023). *Ranking de medios*. https://www.revistapym.com.co/articulos/etiquetados/ranking-medios

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM 2015)*, 399-408. https://doi.org/10.1145/2684822.2685324

Rodríguez, N., & García, F. (2021). Análisis crítico del discurso en medios digitales: estrategias discursivas en contextos de polarización política. *Revista Comunicación y Medios*, *43*, 101-118. https://doi.org/10.5354/0719-1529.2021.65210

Rose-Ackerman, S., & Palifka, B. J. (2016). *Corruption and Government: Causes, Consequences, and Reform* (2.ª ed.). Cambridge University Press.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of Communication*, *49*(1), 103-122.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, *5*(1), 13-22.

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences (PNAS)*, *115*(11), 2584-2589. https://doi.org/10.1073/pnas.1708290115

Strehl, A., & Ghosh, J. (2003). Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, *3*, 583-617. https://www.jmlr.org/papers/volume3/strehl03a/strehl03a.pdf

Superintendencia Nacional de Salud. (2024). *Informe de Rendición de Cuentas 2023–2024*. https://docs.supersalud.gov.co/PortalWeb/planeacion/InformesGestion/RC%20-%20Informe%20Rendici%C3%B3n%20de%20cuentas%202023-2024.pdf

Tandoc, E. C., Jenkins, J., & Craft, S. (2022). The dark side of news values: How news organizations and journalists prioritize controversy and conflict. *Journalism*, *23*(9), 1910-1928. https://doi.org/10.1177/14648849211062132

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(476), 1566-1581. https://doi.org/10.1198/016214506000000302

Transparencia por Colombia. (2022). *Radiografía de la corrupción en salud en Colombia*. https://transparenciacolombia.org.co/analisis-radiografia-corrupcion-2016-2022/

Transparency International. (2021). Corruption in the health sector. En *Global Corruption Report*. Transparency International.

van Dijk, T. A. (2015). Critical Discourse Analysis. En D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The Handbook of Discourse Analysis* (pp. 466-485). Wiley-Blackwell. https://discourses.org/wp-content/uploads/2022/07/Teun-A.-van-Dijk-2015-Critical-discourse-Analysis.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 5998-6008. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Vian, T. (2019). Corruption in the Health Sector. En M. J. Heymann, S. Rushton, & M. Kaldor (Eds.), *The Oxford Handbook of Global Health Politics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190456818.013.25

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, *11*, 2837-2854. https://www.jmlr.org/papers/v11/vinh10a.html

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems (NeurIPS)*, *22*, 1973-1981. https://proceedings.neurips.cc/paper_files/paper/2009/file/385f1ca0c8e8dd1a7b0a17a7b8fa6f7c-Paper.pdf

Yin, H., Sun, Y., Wang, Z., Zhou, Y., & Zhang, C. (2021). Natural language processing for social media data: A review. *Information Fusion*, *64*, 285-303. https://doi.org/10.1016/j.inffus.2020.07.001

Zhang, Y., Wang, H., & Chen, L. (2023). Automated analysis of political corruption discourse using transformer-based language models. *Government Information Quarterly*, *40*(3), 101853. https://doi.org/10.1016/j.giq.2023.101853

12. A. Anexo. Análisis Bibliométrico

El presente análisis bibliométrico se desarrolló a partir de registros recuperados en la base de datos Scopus y del software VOSviewer (versión 1.6.20) para la visualización de redes científicas. La consulta, descarga y procesamiento de la información se realizaron el 22 de febrero de 2026, fecha establecida como punto de corte del conjunto analizado.

Scopus se empleó como fuente principal para la recuperación sistemática de publicaciones correspondientes al periodo 2016–2026, mediante estrategias de búsqueda alineadas con los ejes conceptuales de la investigación. A partir de estas consultas se extrajeron metadatos relacionados con tipo de documento, año de publicación, país de afiliación, área temática, autores y palabras clave, los cuales fueron organizados y depurados para su análisis descriptivo y su representación gráfica.

El análisis se estructuró en torno a dos ejes principales: (i) la producción científica en análisis del discurso y (ii) la producción científica sobre discurso, medios digitales y fenómenos de corrupción. Esta delimitación permitió examinar tanto la consolidación general del campo como el desarrollo de una línea temática específica y emergente.

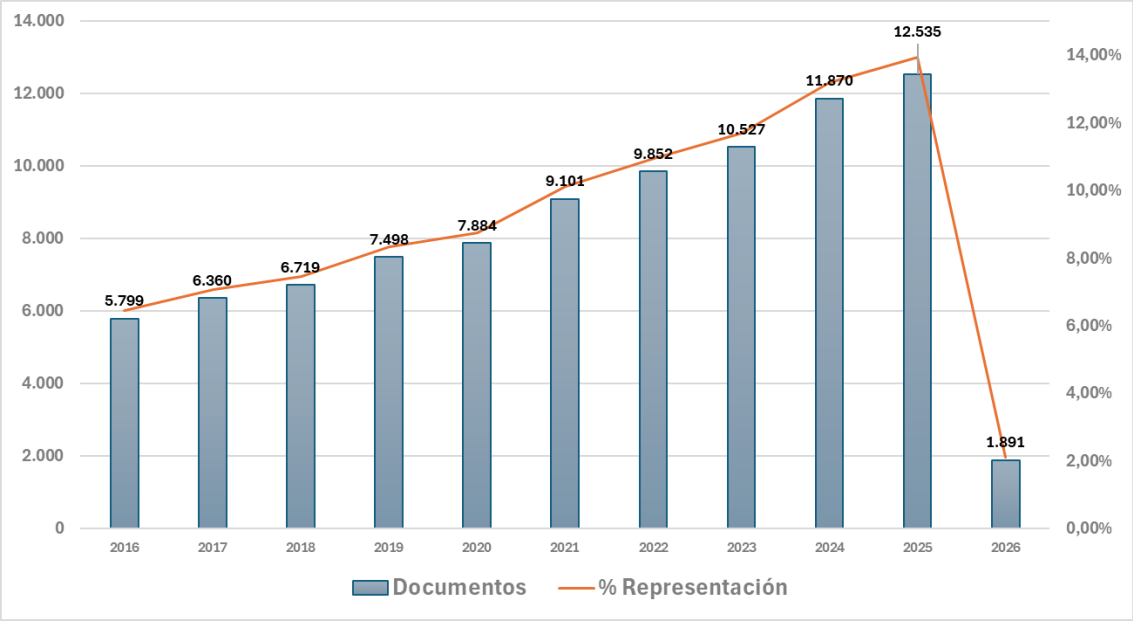
Para la visualización de estructuras temáticas se utilizó VOSviewer, mediante redes de coocurrencia de palabras clave. Las representaciones generales consideran el periodo completo 2016–2026, mientras que el mapa de coocurrencia focalizado se elaboró únicamente con publicaciones de 2025 y 2026, con el fin de identificar tendencias recientes en la producción científica.

12.1. Producción científica en análisis del discurso (2016–2026)

Se utilizó la base de datos Scopus mediante la consulta TITLE-ABS-KEY (“speech” AND “analysis”). La búsqueda arrojó un total de 90.036 documentos para el periodo analizado.

La consulta empleada recupera literatura interdisciplinaria vinculada al estudio del habla y del discurso en sentido amplio, incluyendo enfoques biomédicos, psicolingüísticos y computacionales. En consecuencia, los resultados deben interpretarse como un panorama general de la producción científica relacionada con el análisis del habla y del lenguaje, y no exclusivamente como estudios de análisis crítico del discurso mediático. Este enfoque permite dimensionar la amplitud y la evolución del campo en el que se inscribe la presente investigación.

Figura 46. Distribución anual de publicaciones sobre “speech AND analysis” (2016–2026)



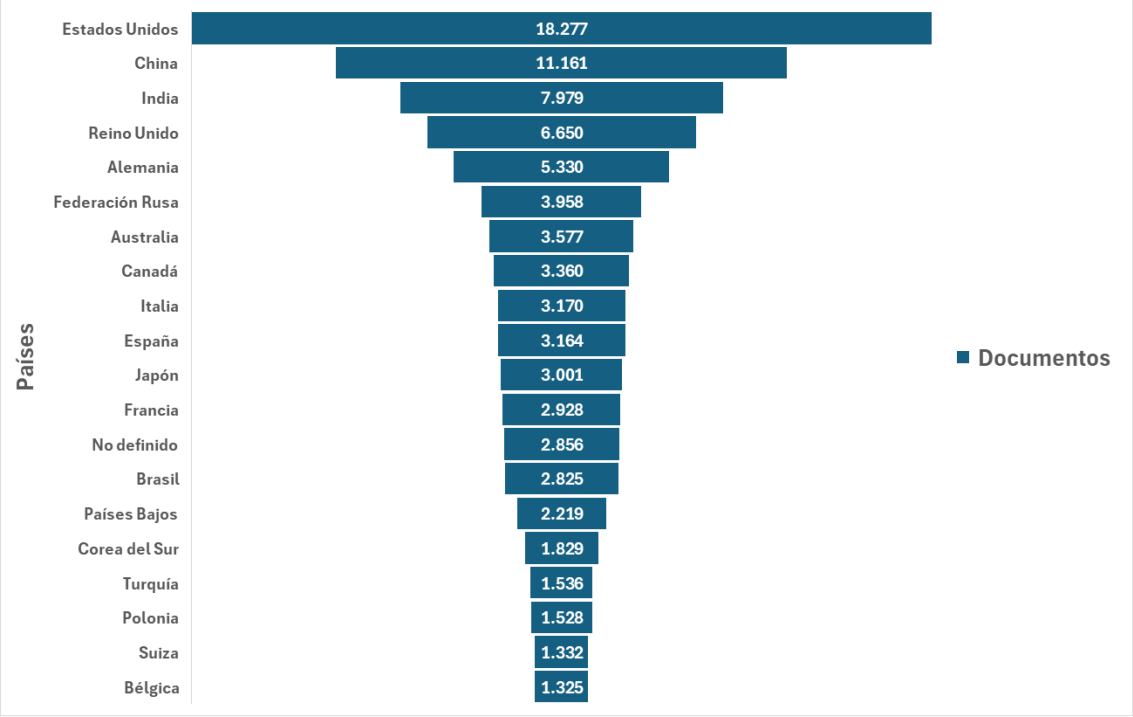
Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“speech” AND “analysis”), realizada el 22 / 02 / 2026.

La evolución temporal muestra un crecimiento sostenido en la producción académica desde 2016 (5.799 documentos) hasta 2025 (12.535 documentos), evidenciando una tendencia ascendente en el interés científico por el análisis del habla y del lenguaje. El incremento es progresivo y consistente a lo largo del periodo, con especial aceleración a partir de 2021.

El valor correspondiente a 2026 (1.891 documentos) debe interpretarse como parcial, dado que la consulta se realizó en febrero de ese año. En consecuencia, no representa una disminución estructural de la producción, sino un corte temporal anticipado.

En términos generales, la tendencia confirma la consolidación del campo y su expansión sostenida durante la última década.

Figura 47. Distribución por país de afiliación “speech AND analysis” (2016–2026)



Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“speech” AND “analysis”), realizada el 22 / 02 / 2026.

La distribución geográfica de la producción científica sobre “speech AND analysis” evidencia una concentración clara en países con alta capacidad investigativa. Estados Unidos lidera ampliamente el volumen de publicaciones, seguido por China e India, mientras que Reino Unido y Alemania completan el grupo de mayor producción. Estos cinco países constituyen el núcleo dominante del campo durante el periodo 2016–2026.

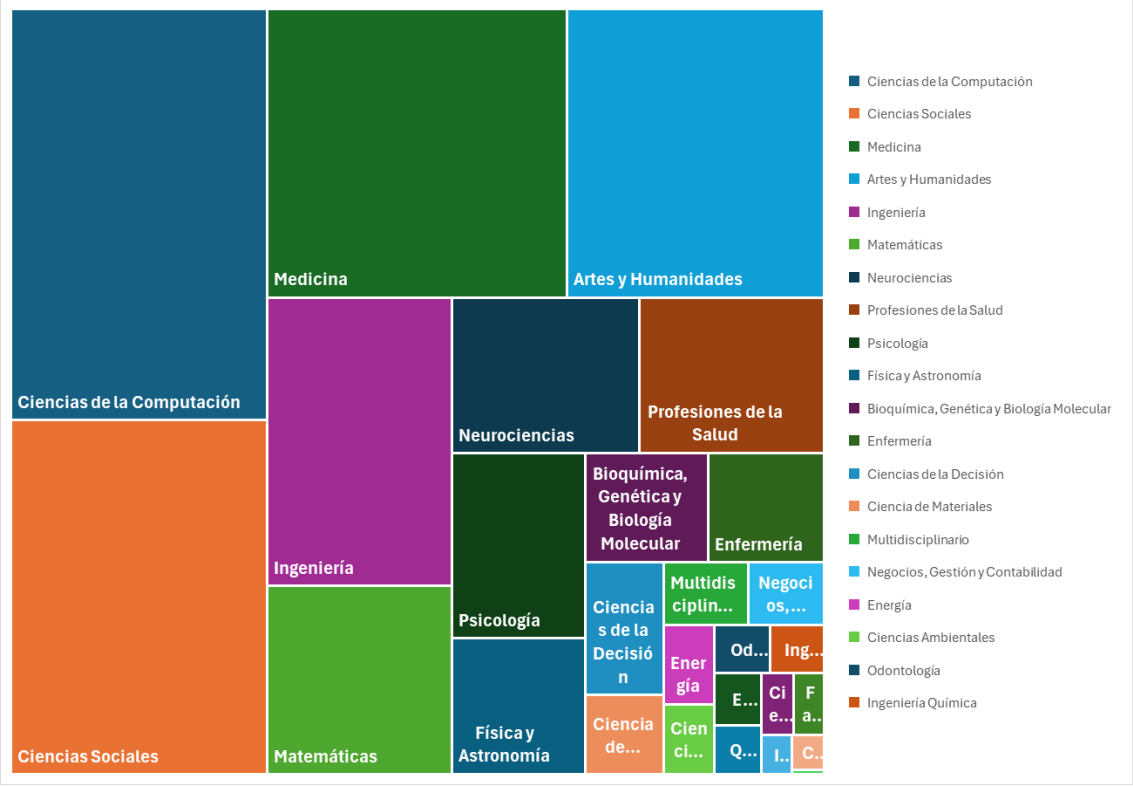
En un segundo nivel se ubican países como la Federación Rusa, Australia, Canadá, Italia y España, junto con Japón y Francia, que mantienen una participación consolidada.

También se observa presencia relevante de Brasil y Países Bajos, así como de otras economías europeas y asiáticas con menor volumen relativo.

Los veinte países representados en la figura concentran en conjunto el 95,5% del total de documentos registrados en el periodo 2016–2026 (86.005 de 90.036 documentos), lo que evidencia una alta concentración geográfica de la producción científica en el campo. El 4,5% restante se distribuye entre otros países con menor volumen de publicaciones, no incluidos en la visualización por criterio de jerarquización descendente.

Esta distribución indica una concentración significativa de la producción científica en economías con sistemas consolidados de investigación en América del Norte, Europa y Asia. No obstante, los datos corresponden exclusivamente a registros indexados en Scopus y deben interpretarse en ese marco de cobertura.

Figura 48. Distribución por área temática “speech AND analysis” (2016–2026)



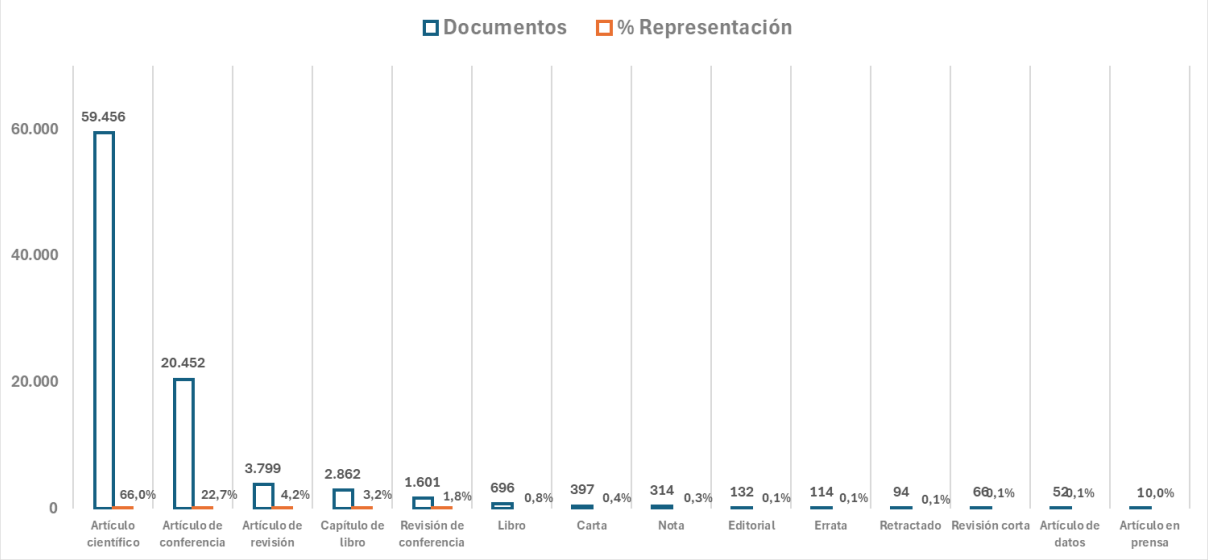
Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“speech” AND “analysis”), realizada el 22 / 02 / 2026.

La distribución por área temática evidencia una marcada interdisciplinariedad en la producción científica sobre “speech AND analysis”. Las mayores concentraciones se observan en Ciencias de la Computación, Ciencias Sociales y Medicina, seguidas por Artes y Humanidades e Ingeniería. Este núcleo disciplinar concentra el mayor volumen de publicaciones del periodo analizado.

En un segundo nivel se ubican Matemáticas, Neurociencias, Psicología, Física y Astronomía, así como Profesiones de la Salud y Bioquímica, Genética y Biología Molecular. La presencia simultánea de áreas técnicas, biomédicas y sociales indica que el análisis del discurso no se restringe a una tradición teórica específica, sino que se integra en múltiples campos de aplicación.

En términos generales, la distribución temática confirma que el campo combina enfoques computacionales, sociales y biomédicos, reflejando tanto su consolidación académica como su expansión hacia dominios interdisciplinarios.

Figura 49. Distribución por tipo de documento “speech AND analysis” (2016–2026)

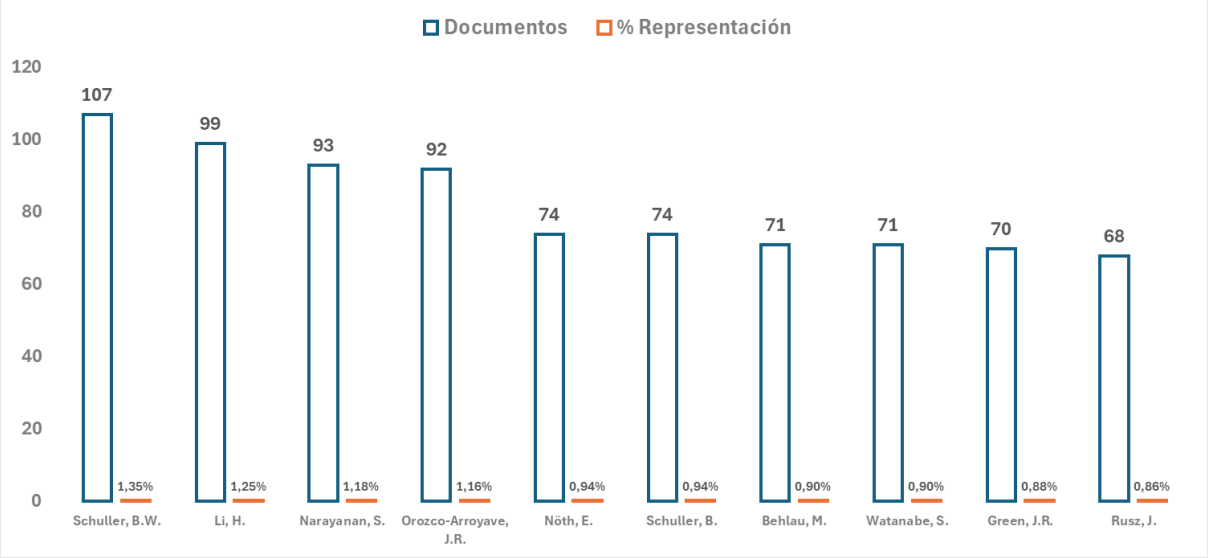


Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“speech” AND “analysis”), realizada el 22 / 02 / 2026.

La estructura documental evidencia que la producción científica en análisis del discurso se desarrolla principalmente a través de artículos académicos, lo que confirma la consolidación del campo en revistas indexadas. La participación relevante de artículos de conferencia indica además una dinámica activa de discusión en eventos especializados, característica de áreas con interacción interdisciplinaria.

Los documentos de revisión y los capítulos de libro ocupan un espacio complementario, asociado a síntesis teóricas y desarrollos conceptuales, mientras que otros formatos presentan una incidencia marginal. En términos generales, la configuración sugiere un campo maduro, con fuerte orientación hacia publicación formal y validación por pares, coherente con su expansión sostenida durante el periodo analizado.

Figura 50. Distribución por autor de publicaciones “speech AND analysis” (2016–2026)



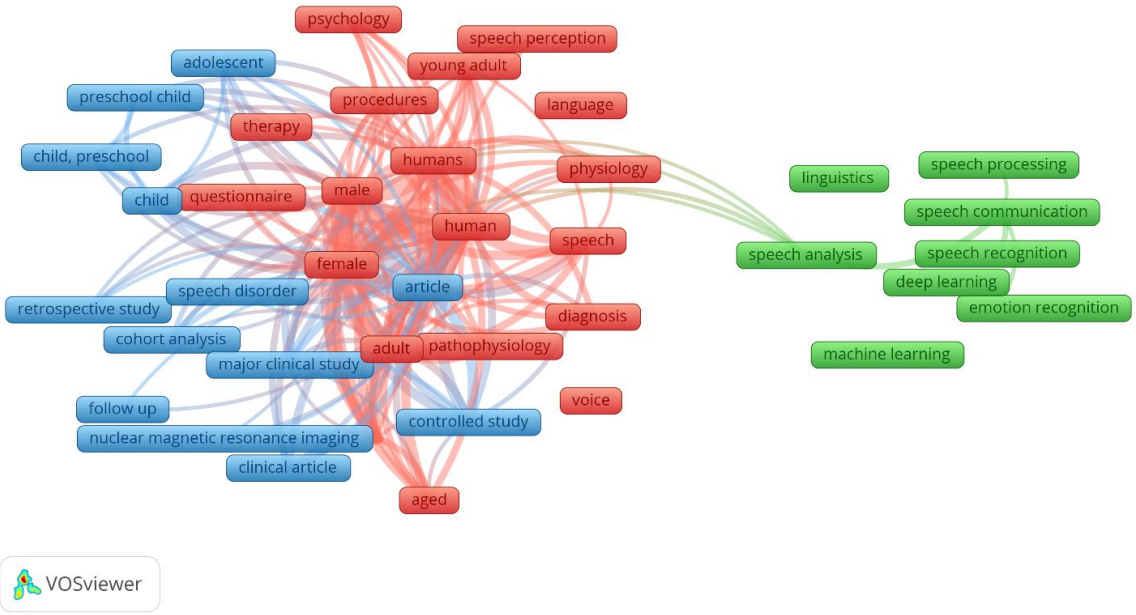
Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“speech” AND “analysis”), realizada el 22 / 02 / 2026.

La distribución por autor evidencia que, si bien existen investigadores con mayor volumen de publicaciones, el campo no se encuentra monopolizado por una sola figura dominante. Los autores con mayor producción presentan diferencias moderadas entre sí, lo que sugiere una estructura académica relativamente distribuida.

La distribución confirma que el liderazgo académico se encuentra compartido entre distintos investigadores, lo cual es consistente con la naturaleza interdisciplinaria y expansiva del área durante el periodo analizado.

Como acción complementaria, se aplicó un filtro temporal restringiendo los resultados a los años 2025 y 2026, obteniéndose 14.531 registros. Sobre este subconjunto se realizó un análisis de coocurrencia de palabras clave para identificar tendencias recientes.

Figura 51. Principales palabras clave 2025–2026 “speech AND analysis”



Nota. Elaborado con base en datos de Scopus mediante la búsqueda TITLE-ABS-KEY (“speech” AND “analysis”), realizada el 22 / 02 / 2026. Visualización generada con VOSviewer 1.6.20.

La red temática evidencia tres agrupamientos principales. El primero se concentra en términos clínicos y biomédicos, asociados a categorías poblacionales y diagnósticas (por ejemplo, “human”, “male”, “female”, “diagnosis”, “physiology”, “pathophysiology”), lo que indica una fuerte presencia de investigaciones en contextos de salud. Un segundo bloque agrupa términos metodológicos y poblacionales relacionados con estudios clínicos y análisis de

cohortes. Finalmente, un tercer clúster, más claramente delimitado, reúne conceptos vinculados a enfoques computacionales como “speech processing”, “speech recognition”, “machine learning”, “deep learning” y “emotion recognition”.

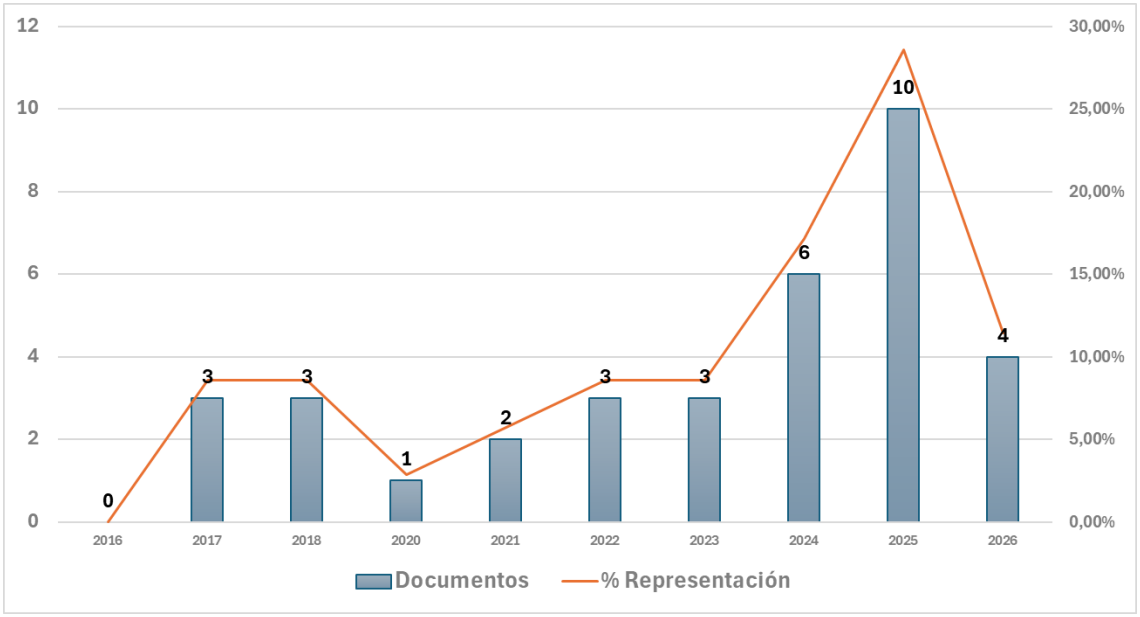
La configuración de la red sugiere que, en el periodo más reciente, el campo articula de manera simultánea enfoques biomédicos y técnicas de aprendizaje automático aplicadas al procesamiento del habla. Esta convergencia es consistente con una orientación interdisciplinaria sostenida, donde la investigación clínica y los métodos computacionales coexisten como ejes relevantes de la producción científica reciente.

12.2. Producción científica discurso en medios digitales y corrupción (2016–2026)

Se utilizó la base de datos Scopus mediante la consulta TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 de febrero de 2026. Los registros obtenidos fueron analizados en función de su distribución temporal, con el fin de identificar patrones de evolución reciente.

La estrategia de búsqueda se orientó a recuperar publicaciones en las que los conceptos de medios, discurso, digitalización y corrupción estuvieran presentes de manera conjunta en título, resumen o palabras clave. Dado el carácter amplio de los operadores booleanos, los resultados incluyen enfoques diversos dentro de las ciencias sociales y la comunicación, por lo que deben interpretarse como un panorama general de la producción académica relacionada con esta intersección temática, más que como un corpus estrictamente homogéneo de estudios sobre análisis crítico del discurso mediático.

Figura 52. Distribución anual de publicaciones “media AND discourse AND digital AND corruption” (2016–2026)

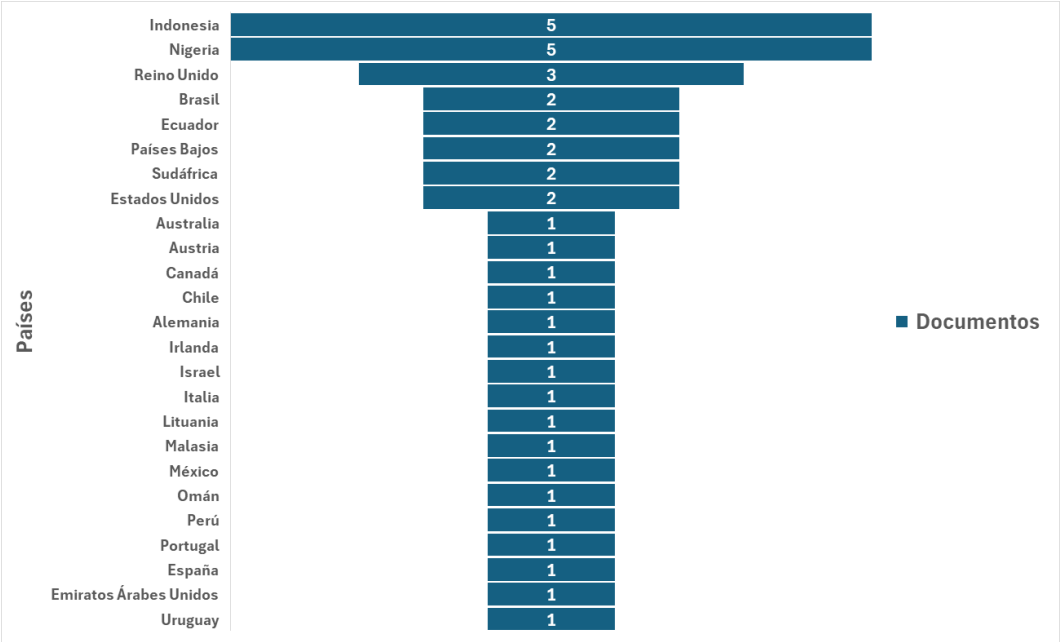


Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 / 02 / 2026.

La distribución temporal evidencia que esta línea temática presenta una producción limitada en los primeros años del periodo analizado, con incrementos progresivos a partir de 2017 y una aceleración más marcada desde 2022. El punto más alto se registra en 2025, lo que confirma una intensificación reciente del interés académico en la articulación entre discurso, medios digitales y corrupción.

El comportamiento observado indica que se trata de una línea emergente dentro del campo más amplio del análisis del discurso. La concentración de publicaciones en los años más recientes sugiere una consolidación progresiva del tema en la agenda científica, especialmente en el contexto de la digitalización de la comunicación pública y el análisis de narrativas asociadas a fenómenos de gobernanza.

Figura 53. Distribución por país de afiliación “media AND discourse AND digital AND corruption” (2016–2026)



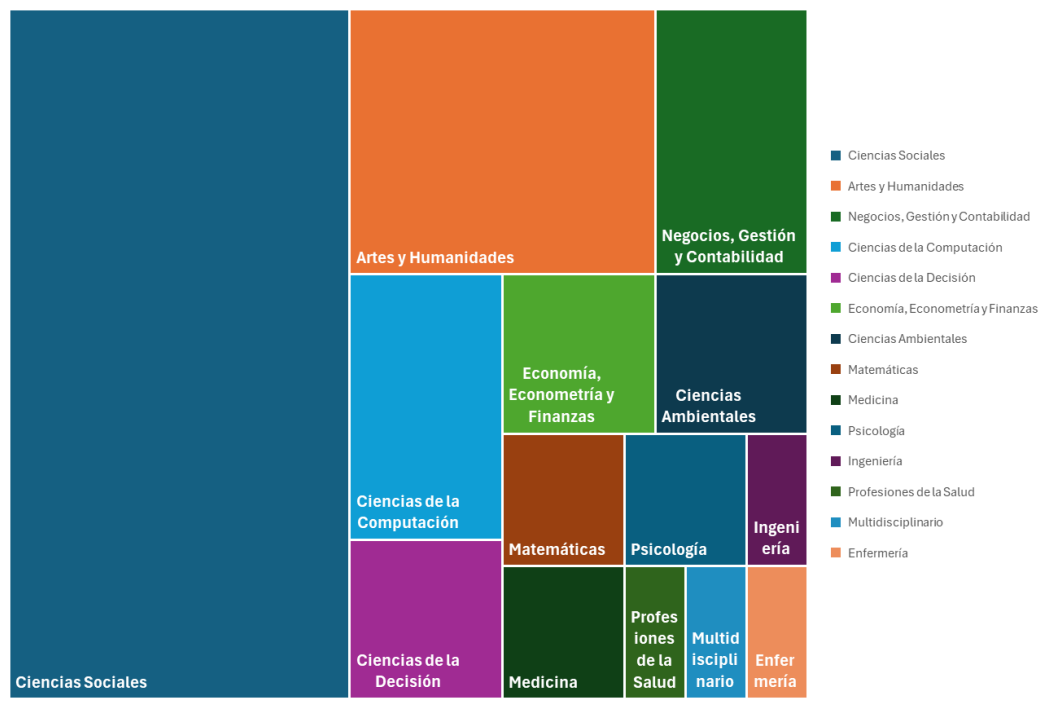
Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 / 02 / 2026.

La distribución geográfica de la producción científica en esta línea temática muestra un patrón altamente descentralizado. Ningún país concentra un volumen dominante de publicaciones, lo que contrasta con la estructura observada en el campo general de “speech AND analysis”.

Indonesia y Nigeria encabezan el listado con el mayor número de documentos, seguidos por el Reino Unido. Este patrón sugiere que la investigación sobre corrupción y medios digitales presenta una asociación con contextos donde los debates sobre gobernanza y transparencia adquieren relevancia pública.

En términos estructurales, el comportamiento geográfico confirma el carácter emergente y transnacional de la línea temática, donde la producción se encuentra distribuida entre diversos contextos nacionales sin consolidación aún de polos académicos predominantes.

Figura 54. Distribución por área temática “media AND discourse AND digital AND corruption” (2016–2026)



Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 / 02 / 2026.

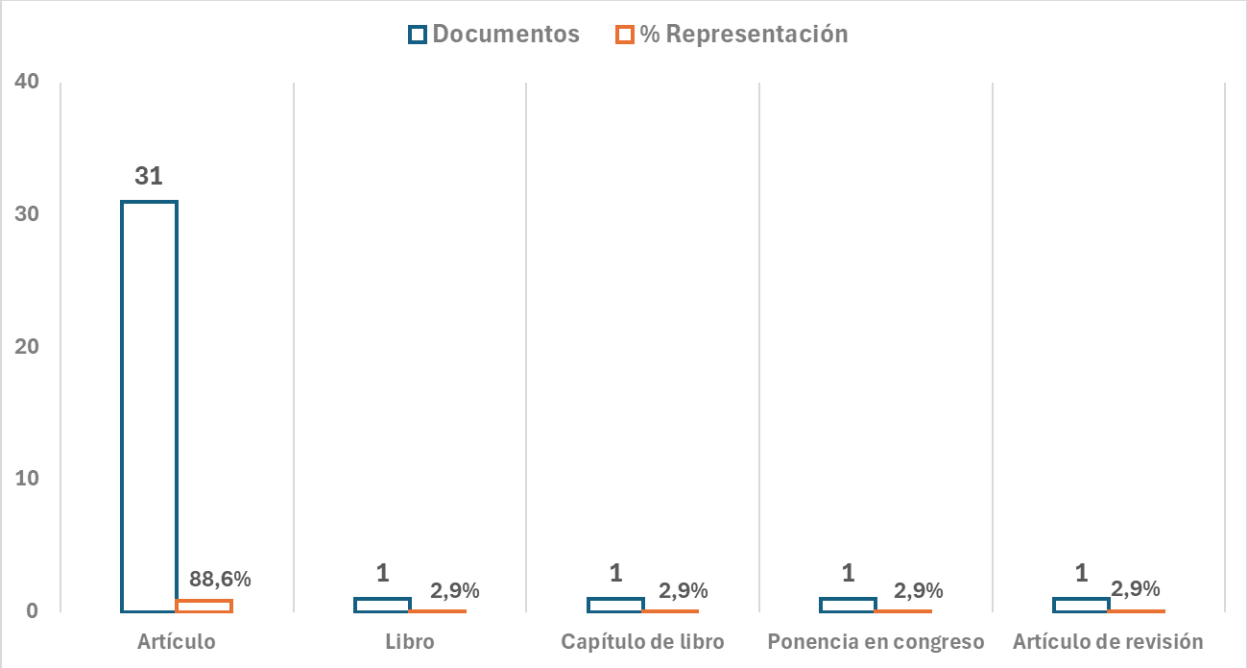
La distribución por área temática muestra un claro predominio de las Ciencias Sociales, que constituyen el núcleo disciplinar del campo. En segundo lugar, se ubican Artes y Humanidades, lo que confirma la fuerte base teórica y crítica desde la cual se aborda el análisis del discurso en contextos mediáticos.

De manera complementaria, se observa participación de áreas como Ciencias de la Computación, Economía y Ciencias de la Decisión, lo que evidencia una apertura hacia enfoques metodológicos y analíticos más técnicos. Asimismo, disciplinas aplicadas como Psicología, Ciencias Ambientales, Medicina y Negocios aportan contribuciones puntuales, reforzando el carácter multidisciplinario de la temática.

En términos estructurales, la configuración confirma que el estudio del discurso en medios digitales sobre corrupción se sustenta principalmente en tradiciones socio-

comunicativas, pero comienza a integrar perspectivas computacionales y analíticas que amplían su alcance metodológico.

Figura 55. Distribución por tipo de documento publicado “media AND discourse AND digital AND corruption” (2016–2026)

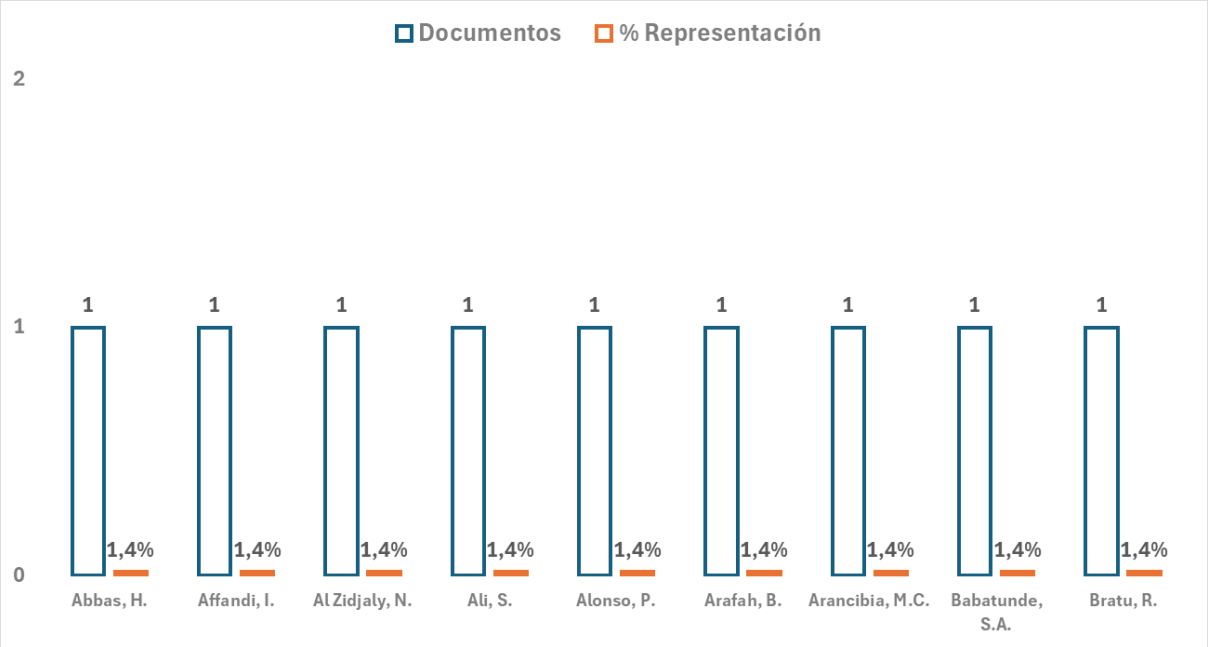


Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 / 02 / 2026.

La estructura documental evidencia una clara predominancia del artículo científico como principal formato de publicación en esta línea temática. Los demás tipos documentales libro, capítulo de libro, ponencia en congreso y artículo de revisión, presentan una participación marginal y aislada.

Esta distribución sugiere que la discusión académica sobre discurso, medios digitales y corrupción se desarrolla fundamentalmente en revistas indexadas, lo que es consistente con una línea de investigación especializada y aún en consolidación. La limitada diversidad de formatos editoriales refuerza la idea de un campo emergente cuya producción se concentra en publicaciones periódicas más que en desarrollos monográficos extensivos.

Figura 56. Distribución por autor de publicaciones “media AND discourse AND digital AND corruption” (2016–2026)



Nota. Elaborado con base en datos extraídos de Scopus mediante la búsqueda TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 / 02 / 2026.

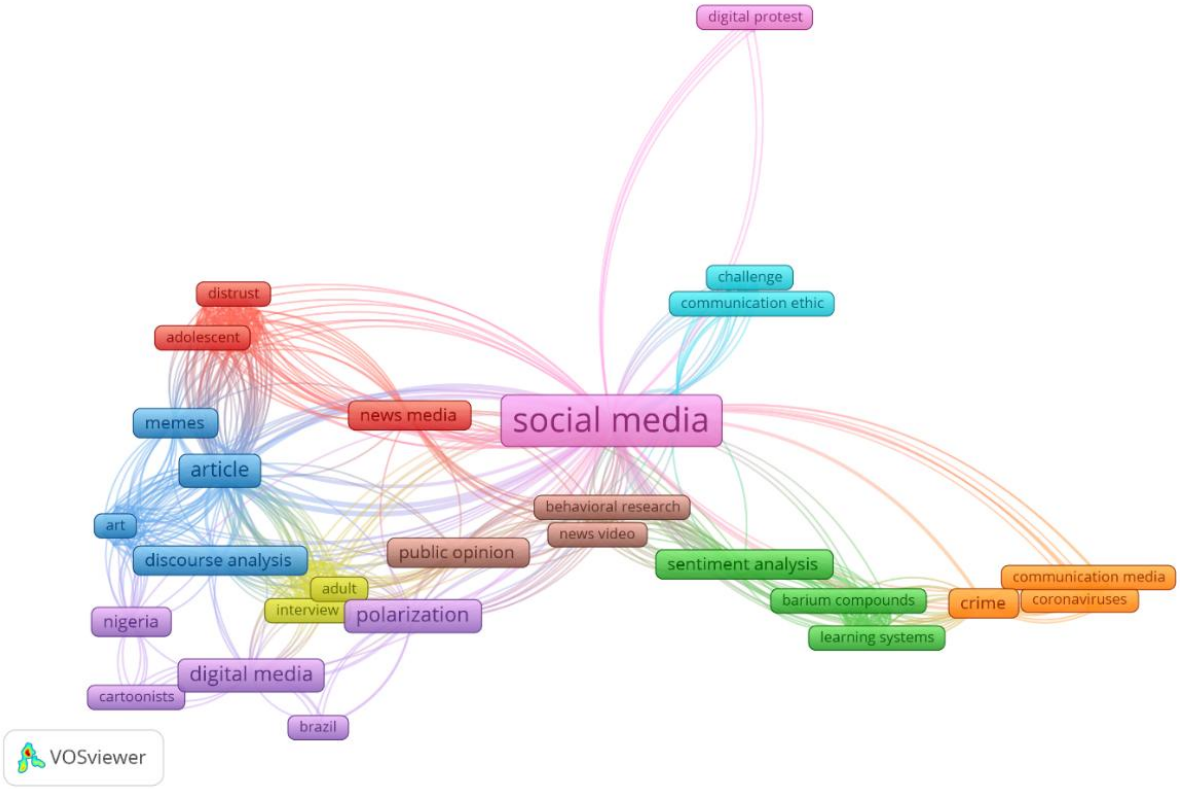
La distribución por autor evidencia una producción altamente dispersa, sin presencia de investigadores con volumen reiterado de publicaciones dentro del periodo analizado. Los autores representados en la figura registran una única contribución cada uno, lo que indica ausencia de concentración o liderazgo académico consolidado en esta línea temática.

Este patrón confirma que el estudio del discurso en medios digitales sobre corrupción se encuentra en una etapa emergente, caracterizada por aportes individuales distribuidos entre distintos investigadores y contextos institucionales. La ausencia de recurrencia autores sugiere que aún no se ha configurado un núcleo estable de producción académica especializado en este cruce temático.

En términos estructurales, la configuración de autores contrasta con la observada en el campo general de “speech AND analysis”, donde sí se identifican investigadores con mayor continuidad productiva.

Como acción complementaria, se aplicó un filtro temporal restringiendo los resultados a los años 2025 y 2026, obteniéndose el subconjunto más reciente de publicaciones. Sobre este conjunto se realizó un análisis de coocurrencia de palabras clave para identificar las relaciones semánticas predominantes.

Figura 57. Principales palabras clave 2025–2026 “media AND discourse AND digital AND corruption”



Nota. Elaborado con base en datos de Scopus mediante la búsqueda TITLE-ABS-KEY (“media” AND “discourse” AND “digital” AND “corruption”), realizada el 22 / 02 / 2026. Visualización generada con VOSviewer 1.6.20 (mapa de coocurrencia por clústeres).

La red semántica evidencia como nodo central el término “social media”, que concentra la mayor densidad de conexiones. A partir de este eje se distinguen agrupamientos temáticos.

Un primer clúster articula términos como “discourse analysis”, “digital media”, “polarization”, “memes” y “news media”, reflejando una orientación hacia el análisis crítico del discurso en entornos digitales y su relación con fenómenos de fragmentación y opinión pública.

Un segundo conjunto se vincula con variables sociales y poblacionales, incluyendo “adolescent”, “distrust” y “public opinion”, lo que sugiere interés en los efectos sociales y percepciones asociadas a contenidos digitales.

Se observa además un bloque conectado con “crime”, “communication media” y “coronaviruses”, que indica la presencia de estudios que relacionan redes sociales con problemáticas específicas de seguridad, desinformación o crisis sanitarias.

Finalmente, términos como “sentiment analysis” y “learning systems” evidencian la incorporación de enfoques analíticos automatizados en el estudio del discurso digital, aunque sin constituir aún el núcleo dominante del campo.

En conjunto, la red confirma que la investigación reciente se estructura alrededor del análisis del discurso en redes sociales, integrando perspectivas críticas, fenómenos de polarización y aproximaciones metodológicas digitales, en un campo aún en consolidación.

12.3. Ampliación exploratoria regional en SciELO

Como complemento al análisis bibliométrico principal realizado en Scopus, se efectuó una ampliación exploratoria en SciELO mediante estrategias de búsqueda formuladas en español, orientadas a identificar producción académica latinoamericana vinculada con análisis del discurso, discurso mediático, medios digitales, corrupción y salud. Esta ampliación tuvo como propósito incorporar literatura regional que pudiera no estar suficientemente representada en la base principal y fortalecer la contextualización del estado del arte en español.

Para ello se utilizaron cuatro ecuaciones de búsqueda: “análisis del discurso” AND “medios digitales”, “discurso mediático” AND corrupción, corrupción AND salud, y corrupción AND “sector salud” AND Colombia. En conjunto, la consulta inicial recuperó 83 registros brutos. Tras un proceso de depuración por duplicados y revisión de coincidencias entre ecuaciones, el universo se redujo a 71 títulos únicos, lo que evidencia tanto la complementariedad como el solapamiento parcial entre búsquedas amplias y focalizadas.

La Tabla 22 resume los resultados obtenidos por ecuación de búsqueda. Se observa que la consulta corrupción AND salud fue la que produjo el mayor volumen de resultados, aunque también presentó el mayor nivel de ruido temático y duplicación interna. En contraste, las ecuaciones más específicas como “discurso mediático” AND corrupción y corrupción AND “sector salud” AND Colombia recuperaron menos registros, pero con mayor precisión temática respecto al objeto de estudio.

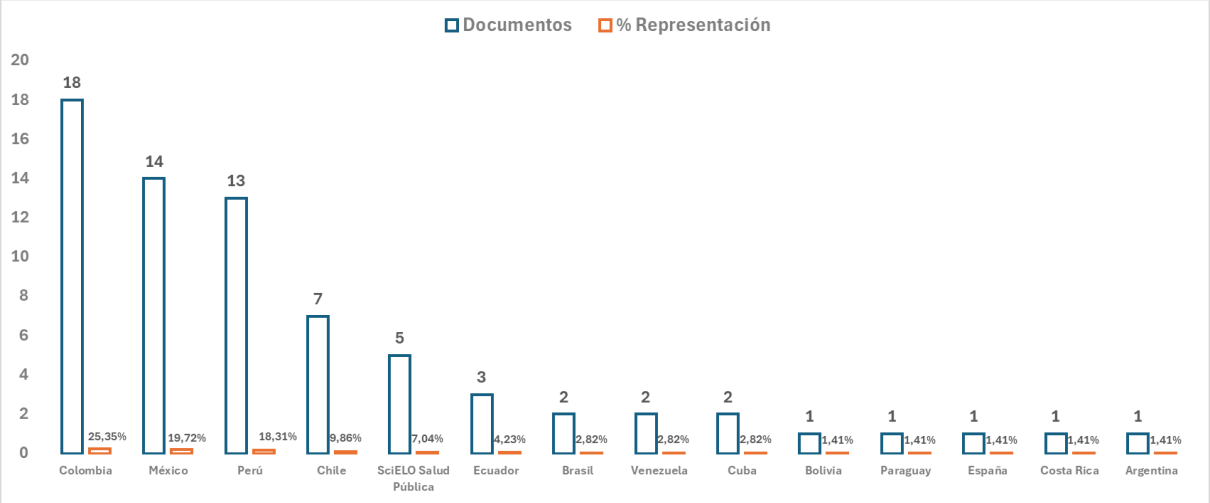
Tabla 22. Resultados de la ampliación exploratoria regional en SciELO según ecuación de búsqueda

Ecuación de búsqueda	Resultados brutos	Propósito analítico	Observación general
“análisis del discurso” AND “medios digitales”	28	Soporte conceptual y metodológico regional sobre discurso y medios digitales	Recupera literatura útil para el marco discursivo y comunicacional, aunque no necesariamente centrada en corrupción o salud
“corrupción” AND “salud”	49	Panorama sustantivo amplio sobre corrupción y salud	Es la ecuación con mayor volumen, pero también la más heterogénea y con mayor ruido temático
“corrupción” AND “sector salud” AND “Colombia”	5	Recuperación focalizada del contexto nacional	Alta especificidad temática y contextual, aunque con baja recuperación
“discurso mediático” AND “corrupción”	1	Soporte discursivo específico	Muy baja recuperación, pero alta precisión frente al objeto analítico

Nota. La tabla resume la ampliación exploratoria regional realizada en SciELO mediante búsquedas en español. Los duplicados removidos corresponden a coincidencias internas y entre ecuaciones, identificadas mediante depuración por título. El total inicial de 83 registros se redujo a 71 títulos únicos tras el proceso de consolidación.

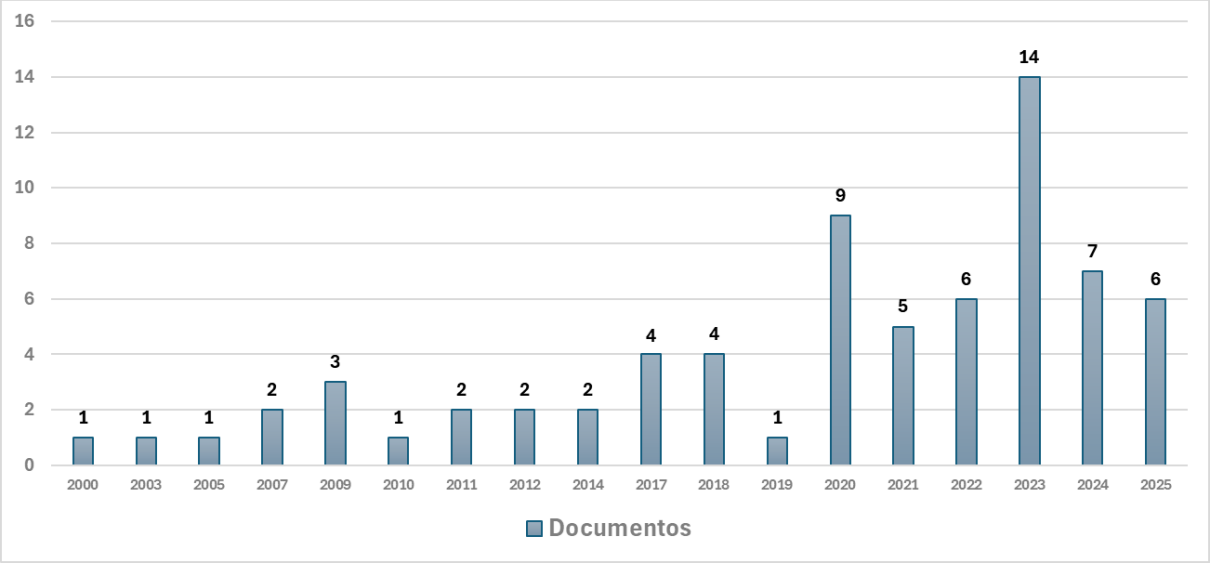
La Figura 58 muestra la distribución de los registros únicos recuperados en SciELO según país o colección de procedencia. Se observa una mayor concentración de resultados en Colombia, México y Perú, con presencia adicional de Chile y de la colección regional SciELO Salud Pública. Esta distribución indica que la ampliación exploratoria en español incorpora producción académica latinoamericana diversa y permite complementar el estado del arte con referencias regionales que no necesariamente aparecen representadas en la misma proporción en la base principal utilizada para el análisis bibliométrico.

Figura 58. Distribución de los registros únicos recuperados en SciELO según país o colección de procedencia



Nota. La figura presenta la distribución de los 71 registros únicos recuperados en SciELO tras la depuración por duplicados. La categoría “SciELO Salud Pública” corresponde a una colección temática regional y no a un país específico.

Figura 59. Distribución temporal de los registros únicos recuperados en SciELO



Nota. La figura presenta la distribución anual de los 71 registros únicos recuperados en SciELO tras la depuración por duplicados. La concentración en años recientes sugiere un crecimiento relativo del interés académico regional en los temas vinculados con discurso, medios digitales, corrupción y salud.

Desde el punto de vista temporal, la producción recuperada muestra una mayor concentración en los años recientes, particularmente entre 2020 y 2025, con un pico en 2023. Este comportamiento sugiere que la literatura regional en español relacionada con medios digitales, corrupción, salud y análisis del discurso presenta un desarrollo más visible en el periodo reciente, aunque con una configuración todavía fragmentaria.

En términos de autoría, la recuperación en SciELO no muestra una concentración sostenida de autores dominantes, lo que sugiere una producción regional todavía fragmentaria y distribuida entre distintos contextos nacionales e institucionales.

En términos sustantivos, la ampliación en SciELO muestra que la literatura regional en español aporta marcos conceptuales y estudios empíricos relevantes sobre análisis del discurso, comunicación política, medios digitales, corrupción y salud pública. Sin embargo, la intersección específica entre corrupción en salud, medios digitales y análisis del discurso sigue siendo reducida y dispersa. En consecuencia, este complemento regional fortalece la contextualización latinoamericana del estado del arte y, al mismo tiempo, refuerza la pertinencia de la presente investigación dentro de un campo que aún no aparece plenamente consolidado en la producción académica regional.

12.4. Alcance analítico y aporte del ejercicio bibliométrico

El análisis bibliométrico desarrollado en este anexo permitió construir una lectura complementaria del estado del arte en dos escalas. Por una parte, la consulta en Scopus ofreció una visión amplia y estructurada de la producción científica internacional asociada al análisis del discurso y, de forma más específica, a la relación entre discurso, medios digitales y corrupción. Esto permitió identificar patrones de crecimiento, distribución geográfica, áreas temáticas dominantes, tipos documentales predominantes y tendencias recientes en la producción académica.

Por otra parte, la ampliación exploratoria en SciELO incorporó una escala regional en español, útil para recuperar literatura latinoamericana vinculada con análisis del discurso, medios digitales, corrupción y salud. La combinación de ambas fuentes fortaleció la contextualización del estado del arte y permitió contrastar la estructura general observada en Scopus con una producción más cercana al contexto lingüístico, temático y geográfico de la investigación.

Desde el punto de vista metodológico, esta estrategia hizo posible diferenciar entre un campo general ya consolidado y una línea temática específica todavía emergente. Mientras el análisis amplio evidenció una producción extensa e interdisciplinaria, la revisión focalizada mostró que la convergencia entre corrupción, medios digitales y discurso mantiene una densidad académica menor y una configuración aún dispersa. La ampliación en SciELO reforzó esta lectura al mostrar que, incluso en la producción regional en español, la articulación específica entre corrupción en salud, medios digitales y análisis del discurso sigue siendo limitada.

En conjunto, el ejercicio bibliométrico no solo contextualiza el problema de investigación, sino que también respalda su pertinencia académica y metodológica. Los resultados obtenidos permiten ubicar este estudio en un campo interdisciplinario en expansión, pero todavía en proceso de consolidación, lo que refuerza la relevancia de una aproximación que articula análisis del discurso, técnicas de PLN y validación empírica para examinar la construcción mediática de la corrupción en salud.

13. B. Anexo. Recursos tecnológicos empleados

El proyecto se ejecutó en un entorno de cómputo local configurado para garantizar autonomía, control experimental y reproducibilidad de todas las fases del análisis, desde la preparación del corpus hasta la validación de resultados mediante técnicas de Procesamiento de Lenguaje Natural (PLN).

El procesamiento se realizó principalmente sobre CPU (AMD Ryzen 5 5600X, 6 núcleos, 12 hilos) con 64 GB de memoria RAM DDR4, capacidad suficiente para manejar en memoria el corpus completo y las matrices derivadas del modelado temático. El almacenamiento se efectuó en unidades SSD NVMe, lo que permitió optimizar operaciones de lectura y escritura durante el procesamiento.

Los métodos centrales empleados preprocesamiento lingüístico, modelado LDA, análisis de sentimiento y cálculo de métricas de coherencia están diseñados para ejecución eficiente en CPU, por lo que no se requirió aceleración mediante GPU.

El entorno de desarrollo se gestionó mediante Python y bibliotecas especializadas de PLN, utilizando entornos virtuales aislados y control de versiones mediante GitHub, lo que garantiza la trazabilidad y replicabilidad del pipeline analítico.