Análisis de la cobertura de la Wikipedia en la prensa hispanohablante entre los años 2013 y 2023

Analysis of Wikipedia Coverage in Spanish-Language Media between 2013 to 2023

Boté-Vericad, J. J.

Juan-José Boté-Vericad. Universidad de Barcelona (España)
Doctor en Información y Documentación por la Universidad de Barcelona en España y Doctor en Lingüística y Ciencias de la Información por la Universidad de Hildesheim en Alemania. Es profesor del departamento de Biblioteconomía, Documentación y Comunicación Audiovisual en la Universidad de Barcelona. Sus líneas de investigación incluyen el comportamiento ante la información y la ciencia abierta.
https://orcid.org/0000-0001-9815-6190, juanjo.botev@ub.edu

Recibido: 24-09-2024 – Aceptado: 19-01-2025
https://doi.org/10.26441/RC24.1-2025-3726

RESUMEN: Este artículo ofrece un análisis de la cobertura de Wikipedia en las noticias de medios digitales hispanohablantes. Se aplica la Teoría del Encuadre para examinar cómo los medios de comunicación presentan Wikipedia en los titulares de sus artículos. Se analizan 652 noticias extraídas de la base de datos Factiva entre los años 2013 y 2023. Se realizan diferentes análisis como la distribución y la tendencia temporal de las noticias, la frecuencia y mapa de calor de palabras, el algoritmo de asignación latente de Dirichlet (LDA), la co-ocurrencia de palabras en el contenido y títulos, aplicando procesamiento de lenguaje natural y técnicas de *machine learning* para el análisis de temas. El resultado del análisis permite observar que la prensa española es la que más ha publicado sobre Wikipedia, con un aumento en la cobertura durante eventos globales, como la pandemia de COVID-19 y el conflicto en Ucrania. También se observan controversias en torno a biografías de figuras públicas, especialmente políticos, en momentos clave. Además, el análisis revela un sesgo de género, ya que las mujeres participan menos en la edición de Wikipedia, y el contenido relacionado con ellas es eliminado con mayor frecuencia. Se concluye que es necesario promover una mayor diversidad en la comunidad de personas editoras y realizar más acciones para mitigar los sesgos en la plataforma.

Palabras clave: enciclopedias; prensa y lengua española; cobertura informática; teoría de la comunicación; análisis de datos; análisis de tendencias; análisis de contenido; comunidades virtuales; igualdad de género; análisis léxico.

ABSTRACT: This article analyses Wikipedia's coverage in news from Spanish-speaking digital media. Framing Theory is used to examine how media outlets present Wikipedia in their article headlines. A total of 652 news articles were analyzed from the Factiva database between the years 2013 and 2023. Various analyses were conducted, including the distribution and temporal trends of the news, word frequency and heatmaps, the Latent Dirichlet Allocation (LDA) algorithm, and word co-occurrence in content and headlines. Natural language processing and machine learning techniques were applied for topic analysis. The results show that Spanish media has published the most about Wikipedia, with increased coverage during global events such as the COVID-19 pandemic and the Ukraine conflict. Controversies related to the biographies of public figures, particularly politicians, are also highlighted during key moments. Furthermore, the analysis reveals a gender bias, with women participating less in Wikipedia editing and content related to them being more frequently deleted. The study concludes that there is a need to promote greater diversity within the editing community and to implement further measures to mitigate biases on the platform.

Keywords: encyclopedias; press and spanish language; information coverage; communication theory; data analysis; trend analysis; content analysis; virtual communities; gender equality; lexical analysis.

1. Introducción

Este artículo analiza cómo la prensa hispanohablante cubrió Wikipedia entre los años 2013 y 2023, examinando las temáticas destacadas y las tendencias temporales en las noticias digitales de España y Latinoamérica. En esta cobertura mediática, los encuadres noticiosos desempeñan un papel crucial en la forma en que Wikipedia es presentada al público. La Teoría del Encuadre (Framing Theory), desarrollada y aplicada por diversos autores (Muñiz, 2011; Piñeiro-Naval y Mangana, 2018; Zheng, 2020; Casado-Gutiérrez, 2021), resulta útil para entender cómo los medios seleccionan y enfatizan ciertos aspectos de Wikipedia, moldeando la percepción pública y el debate en torno a la enciclopedia libre. Esta teoría permite explorar los posibles sesgos y la influencia que los medios ejercen sobre la audiencia al cubrir temas relacionados con Wikipedia, reflejando y amplificando determinados intereses y preocupaciones.

La Wikipedia en español abarca todos los países de habla hispana. Por ello, la versión en español de Wikipedia cuenta con diversas comunidades que generan contenido, además de las personas editoras que contribuyen de manera individual. Las normas en Wikipedia están basadas en directrices de publicación (Yang y Colavizza, 2024; Petroni et al., 2023). No obstante, la enciclopedia libre tiene la ventaja de permitir que cualquier persona edite el contenido, que luego es debatido por distintas personas editoras. Sin embargo, la aprobación o eliminación de contenido no está exenta de posibles sesgos (Ferran-Ferrer et al., 2023; Kaffee et al., 2023; Morris-O'Connor et al., 2022; Gluza et al., 2021).

Wikipedia permite ediciones abiertas, creando una agenda cuasi-pública donde usuarios agregan o revisan contenido, reflejando sus intereses (Lee, 2018). Aun así, las pequeñas contribuciones de las grandes masas son necesarias en el sistema debido a su utilidad al aportar nuevas perspectivas a los artículos (Chhabra y Iyengar, 2020). En Wikipedia se encuentran contenidos de todo tipo, aunque algunos estudios confirman que hay sesgos de publicación dado que las mujeres tienden a publicar mucho menos contenido que los hombres (Hinnosaar, 2019; Ferran-Ferrer et al., 2023). También muchos contenidos de mujeres se borran a una velocidad mucho más alta que los contenidos generados por hombres (Morris-O'Connor et al., 2022; Ferran-Ferrer et al., 2023). Esto provoca como consecuencia, sesgos, discriminación y perpetuidad de estereotipos masculinos en la propia enciclopedia.

La Wikipedia por su parte se utiliza en la prensa escrita como fuente de información y también como fuente de difusión o establecimiento de la agenda. Messner y South (2011) analizaron cómo los periódicos en Estados Unidos encuadran a Wikipedia, destacando tanto su utilidad como las dudas sobre su fiabilidad. Ren y Xu (2024) complementan esta visión, mostrando cómo Wikipedia interactúa con las agendas mediáticas globales, reflejando narrativas políticas y culturales internacionales. Estos estudios resaltan la importancia de comprender cómo los medios configuran la percepción pública de Wikipedia en contextos diversos. Así Debus y Florczak (2022) realizaron un estudio sobre los comunicados de prensa y datos de vistas de páginas de Wikipedia para estudiar la dinámica de temas de un partido populista de derecha, la AfD en Alemania. De 2013 a 2019, analizaron las visitas a Wikipedia y reflejaron que la atención sobre la política económica disminuye mientras aumenta en la migración (Carmel, 2013). Por su parte, Lee (2018) estudió la dinámica del establecimiento de la agenda entre cinco medios de comunicación importantes del Reino Unido y Wikipedia, considerándola una forma de periodismo participativo. El objetivo fue evaluar cómo se seleccionan los marcos y los sentimientos en artículos sobre el Brexit (Hobolt et al., 2021; Walter, 2019) y en las páginas de Wikipedia. Utilizó técnicas de minería de textos para analizar los marcos y sentimientos de los artículos de noticias y las ediciones de Wikipedia desde la creación de la página "Brexit" hasta el referéndum. La metodología incluyó clasificación automática de textos y análisis de sentimientos basado en léxico, así como pruebas de causalidad de Granger para explorar relaciones entre las agendas. Los resultados mostraron que los marcos y sentimientos variaban entre los medios y Wikipedia, reflejando diferentes intereses y enfoques. Concluyó que Wikipedia, al permitir

la participación ciudadana, ofrece una perspectiva valiosa sobre los temas de interés público y complementa el análisis de los medios tradicionales. Silva y Barbosa (2022) mejoraron la comprensión de noticias digitales al correlacionar artículos con tablas web, como las de Wikipedia, usando un modelo de atención basado en BERT, el modelo de procesamiento de lenguaje natural desarrollado por Google. Su modelo superó técnicas estándar de Recuperación de Información en precisión y ranking recíproco medio (MRR), demostrando su eficacia en augmentación de noticias. El *Mean Reciprocal Rank* (MRR) es una métrica que evalúa la eficacia de un sistema de búsqueda, calculando la media del valor recíproco de la posición de la primera respuesta relevante en una lista de resultados. Cuanto mayor es el MRR, más cerca del inicio aparecen las respuestas relevantes.

Así pues, esto nos lleva a las siguientes preguntas de investigación:

- PI: ¿Cuáles son los términos, conceptos y temáticas destacadas en la prensa digital hispanohablante sobre la Wikipedia?
- PI2: ¿Qué relación hay entre los medios digitales y el tipo de artículo que publican?
- PI3: ¿Qué relaciones y co-ocurrencias de palabras se pueden identificar en los títulos y contenidos de los artículos sobre Wikipedia?

2. Revisión de la literatura

2.1. La Teoría del Encuadre

La Teoría del Encuadre se centra en cómo los medios de comunicación estructuran y presentan la información, afectando la interpretación del público. Muñiz (2011) analizó cómo los periódicos digitales mexicanos encuadran las noticias sobre migración utilizando la Teoría del Encuadre. A través de un análisis de contenido de 228 noticias publicadas en 2007, identificó cuatro encuadres principales: «debate político sobre regulación migratoria», «delincuencia y expulsión de migrantes», «procesos de regularización» y «experiencia migratoria». El estudio reveló que la mayoría de las noticias sobre migración se centraban en hechos ocurridos en Estados Unidos, mientras que la cobertura de eventos en México, especialmente en la frontera sur, era escasa. Este trabajo destaca la importancia de los encuadres noticiosos en la formación de percepciones públicas sobre la migración. Piñeiro-Naval y Mangana (2018) realizaron una revisión de la literatura sobre la Teoría del Encuadre, enfocándose en su evolución en el contexto hispanoamericano durante los años 2007 a 2016. Para ello, analizaron las 10 revistas de mayor impacto en 2016. Los resultados indicaron un crecimiento constante en la producción científica sobre el tema, destacando la contribución de un pequeño grupo de autores muy activos. Concluyeron que la Teoría del Encuadre ha ganado relevancia en la investigación en comunicación en el mundo hispano, posicionándose como un paradigma clave tanto teórica como metodológicamente. También Piñeiro-Naval et al. (2018) analizaron cómo los municipios españoles divulgan su patrimonio cultural a través de portales web, utilizando la Teoría del Encuadre y la Teoría de la Identidad Social. Mediante un análisis de contenido web, identificaron las estrategias de comunicación utilizadas para promover la identidad colectiva y destacar elementos culturales, tanto materiales como inmateriales. Los resultados revelaron diferencias significativas en la presentación y enfoque del patrimonio según el tamaño del municipio y su ubicación geográfica, subrayando la importancia de la identidad local en la comunicación digital.

Por su parte, Pérez-Salazar (2019) utilizó la Teoría del Encuadre para analizar cómo los memes compartidos en Facebook y Twitter encuadraron la escasez de gasolina en México en enero de 2019. El objetivo fue identificar cómo estos memes contribuyeron a la propagación de marcos interpretativos en torno a este evento mediático. A través de un enfoque cualitativo,

el autor analizó 55 memes, identificando dos tipos principales de encuadres: amplificación y extensión. Los resultados muestran que los memes amplificaron la crítica hacia el gobierno de Andrés Manuel López Obrador y extendieron la discusión hacia temas medioambientales y de organización ciudadana. Concluyó que los memes son herramientas clave para construir significados y fomentar la acción colectiva en la esfera digital. Zheng (2020) examinó el uso de marcos metafóricos y no metafóricos en los informes de los medios estatales chinos sobre la COVID-19 (Mutua y Oloo, 2020), empleando la Teoría del Encuadre y la Teoría de la Metáfora Conceptual. Descubrió que marcos como «guerra» y «solidaridad» se utilizaron para movilizar a la población, fortalecer la cohesión social y aumentar la confianza en las medidas gubernamentales. Estos marcos desempeñaron un papel clave en la comunicación eficaz y en la gestión de la crisis sanitaria.

Así, Johnson et al. (2024) investigaron la evolución de la Política Nacional Oceánica en Estados Unidos, comparando las prioridades de las administraciones de Obama y Trump. Utilizando la Teoría del Encuadre, analizaron cómo se presentaron y priorizaron las políticas oceánicas a lo largo de doce años. Encontraron cinco prioridades comunes: una política oceánica estratégica, administración de ecosistemas, impacto económico, control federal frente al estatal, y participación de las partes interesadas. Sin embargo, las administraciones difirieron en su implementación, con Obama enfocándose en la planificación anticipatoria y Trump en el crecimiento económico. Concluyeron que las políticas por orden ejecutiva carecen de impacto duradero, subrayando la necesidad de una cooperación bipartidista para crear una política oceánica sólida y sostenida.

Aunque la Teoría del Encuadre ha sido ampliamente utilizada, también ha recibido críticas significativas. Valera-Ordaz (2016) señala un «sesgo mediocéntrico» en la aplicación del framing en estudios de comunicación, al centrarse exclusivamente en las prácticas periodísticas y omitir las influencias políticas e ideológicas en la construcción de los marcos. De igual manera, Sádaba (2001) subraya la importancia de abordar las limitaciones metodológicas de los estudios de encuadre, especialmente al aplicar herramientas automatizadas de análisis. Estas críticas resultan relevantes en investigaciones actuales, particularmente en contextos donde los sesgos inherentes a las herramientas utilizadas pueden influir en los resultados, como en el análisis financiero o mediático.

2.2. Minería de textos en el análisis de noticias

La minería de textos se ha consolidado como una herramienta en el análisis de grandes volúmenes de datos textuales, especialmente en el ámbito de las noticias digitales (Pinto et al., 2023; Ptaszek et al., 2024, Wirawan et al., 2024). A través de técnicas de procesamiento de lenguaje natural (PLN) y aprendizaje automático, es posible extraer información, identificar patrones y descubrir temas subyacentes que no son evidentes a simple vista. En este contexto, la minería de textos permite desglosar y comprender mejor los contenidos mediáticos, revelando las temáticas predominantes, los posibles sesgos, encuadres y tendencias que configuran la percepción pública. Este apartado examina el uso de estas técnicas en el análisis de noticias, con un enfoque particular en cómo el PLN y el *machine learning* han sido aplicados para analizar la cobertura mediática de Wikipedia.

2.3. El procesamiento del lenguaje natural en el análisis de noticias

El procesamiento de lenguaje natural en el texto permite realizar análisis como clasificación de noticias, el análisis de sentimientos o la identificación de temas ocultos mediante la asignación latente de Dirichlet. En su estudio, Urologin (2018) propone un enfoque novedoso que combina el resumen de textos y el análisis de sentimientos en artículos de noticias, utilizando el analizador de sentimientos VADER y técnicas de aprendizaje automático como la Regresión Logística, Random Forest y AdaBoost. El estudio se centra en artículos de la BBC sobre deportes,

implementando métodos de resumen basados en la sustitución de pronombres para mejorar la precisión del análisis de sentimientos. Los resultados muestran una alta tasa de clasificación de sentimientos, con una visualización 3D innovadora que ofrece una mejor representación de la información de sentimientos, logrando una clasificación máxima del 84.93% sin resumir, y mejorando con diferentes ratios de resumen hasta un 83.23%. Por su parte, Krishnamoorthy (2018) presenta un modelo jerárquico de clasificación de sentimientos para artículos financieros, utilizando indicadores de rendimiento y minería de reglas de asociación para predecir la polaridad del sentimiento. El modelo mejora la precisión y gestiona de forma eficaz los desequilibrios de clase en textos financieros, beneficiando a inversores y analistas.

Otros autores desarrollaron un algoritmo para detectar sesgos en las noticias mediante análisis de sentimientos. El sistema recopila el contenido de noticias de varios portales en línea, analiza la polaridad del sentimiento de los textos, y determina la parcialidad de las noticias con un puntaje que varía entre -1 y 1. Los resultados experimentales mostraron que el algoritmo mejora la precisión en la identificación de sesgos, ofreciendo una herramienta útil para garantizar la imparcialidad y fiabilidad en la información (Sv y Geetha, 2019). También Boté-Vericad (2022) analiza los perfiles de Spotify en Twitter dirigidos a países de habla hispana, incluyendo Argentina, Chile, Colombia, México y España. El estudio examina cómo las variaciones lingüísticas del español influyen en la interacción y el compromiso de la audiencia. Utilizando análisis de sentimiento, modelado de temas y el uso de hashtags, se concluye que las variaciones lingüísticas impactan significativamente en el compromiso de los usuarios, sugiriendo que las estrategias de marketing deben adaptarse a las diferencias lingüísticas regionales para optimizar la interacción en redes sociales.

Así, Prasad et al. (2023) realizan una revisión sistemática de los métodos de procesamiento de lenguaje natural para la clasificación de sentimientos en artículos de noticias en línea. Analizan modelos como VADER, TextBlob, SVM, RNNs y BERT, destacando su efectividad en la identificación automática de sentimientos. Concluyen que, aunque la elección del modelo depende de las necesidades específicas del proyecto, los modelos basados en *transformers* han demostrado ser especialmente eficaces, ofreciendo herramientas valiosas para la toma de decisiones en diversos campos. Mishra et al. (2023) presentan un modelo basado en procesamiento de lenguaje natural para la automatización del resumen de textos y el análisis de sentimientos en noticias. Utilizando técnicas de segmentación en unidades léxicas (también tokenización) y bibliotecas como NLTK, el modelo primero resume el texto antes de realizar el análisis de sentimientos. Este enfoque mejora la precisión y eficiencia, logrando un 91.67% de exactitud en la clasificación de emociones. Concluyen que el modelo puede ser extendido a otros tipos de grandes volúmenes de datos.

2.4. El uso de *machine learning* en el análisis de temáticas

La Asignación Latente de Dirichlet (LDA en inglés), un modelo de *machine learning*, es comúnmente utilizado en el procesamiento de lenguaje natural para descubrir temas ocultos en textos. Keswani et al. (2020) presentan un modelo basado en LDA para extraer características textuales de artículos financieros y predecir tendencias en el mercado de valores. El modelo propuesto incorpora datos históricos del S&P 500 junto con artículos de noticias financieras, permitiendo una mejor predicción de caídas o subidas del mercado. La investigación destaca que la combinación de LDA con parámetros adicionales específicos del mercado mejora la precisión de las predicciones, sugiriendo su potencial aplicación en modelos de aprendizaje automático más complejos para la previsión financiera. Concluyen que la metodología puede ser expandida para incluir más indicadores económicos, mejorando aún más la capacidad predictiva del modelo.

También Liu (2020) analiza la relación entre noticias destacadas y el mercado de valores utilizando el modelo LDA y el método de estudio de eventos. El estudio clasifica las noticias en «noticias de última hora» y «noticias continuas», encontrando que solo las primeras provocan

rendimientos anormales a corto plazo en el mercado. Concluye que las noticias de última hora tienen un impacto significativo en las reacciones del mercado, mientras que las noticias continuas no afectan significativamente a las decisiones de inversión, ya que la información ya ha sido integrada previamente en los precios de las acciones.

Por su parte, Shao et al. (2022) desarrollan un modelo mejorado de LDA para la clasificación de noticias, introduciendo una nueva representación de texto (Cnew) y un método de iteración adaptativa en el muestreo de Gibbs. El modelo se probó en dos corpus de noticias en chino, mostrando mejoras significativas en la precisión, el recuerdo y la medida F1, en comparación con el modelo LDA tradicional. Concluyen que su enfoque es eficaz para mejorar la clasificación de textos en aplicaciones de procesamiento de lenguaje natural. Yang et al. (2022) presentan un método adaptativo para determinar el número óptimo de temas en el modelo de LDA para la identificación de temas en noticias. Al combinar análisis semántico y de series temporales, el modelo mejorado utiliza el algoritmo Co-DPSC para optimizar la selección del número de temas, logrando una mejora significativa en la precisión y el valor F. Los resultados muestran que el método propuesto supera las técnicas tradicionales, aunque su aplicación puede estar limitada a conjuntos de datos específicos.

3. Metodología

Para la realización de este estudio, se han establecido los siguientes criterios. En primer lugar, se analizaron 10 periódicos digitales de habla hispana de cada país listado en la Tabla 1, de acuerdo con el Scimago Media Ranking 2024 Summer Edition. La selección de estos 10 medios sigue los principios de los estudios de caso (Yin, 2003) tomando en consideración los siguientes supuestos: si algún periódico no arrojaba resultados, se utilizaban hasta un máximo de dos medios adicionales por país. Si aun así no se obtenían resultados, se aplicaba el filtro por región en la base de datos Factiva, lo que permitía seleccionar únicamente el país específico.

Tabla 1. Relación total de noticias por países

Región	País	Noticias
	España	345
México y Centroamérica	México	59
	Guatemala	7
	Costa Rica	0
	El Salvador	0
	Honduras	0
	Nicaragua	0
	Panamá	7
Caribe	Cuba	0
	República dominicana	0
	Puerto Rico	0
Andina	Colombia	18
	Perú	23
	Venezuela	1
	Ecuador	0
	Bolívia	7
Cono Sur	Argentina	88
	Chile	79
	Uruguay	13
	Paraguay	5
Total		**652**

Fuente: Elaboración propia.

En segundo lugar, para la recuperación de noticias se utilizó la base de datos Factiva, que cubre la prensa internacional con más de 20 millones de artículos. Se buscó el término «Wikipedia» en el título de las noticias dentro de una ventana temporal de 11 años completos (2013 a 2023). El objetivo era analizar artículos cuyo argumento principal fuera Wikipedia. En este contexto, se aplicó la Teoría del Encuadre de Goffman (1974). Esta teoría sugiere que la forma en que los medios estructuran y destacan la información influye en la percepción pública. La teoría es particularmente útil para analizar los títulos de las noticias, identificando los marcos o «frames» utilizados para representar Wikipedia y cómo estos pueden influir en la percepción de los lectores. Concretamente, se examinó cómo los medios presentan Wikipedia en los titulares de sus artículos. Todas las noticias fueron revisadas mediante una lectura; aquellas que no tenían relación con Wikipedia o sus comunidades fueron descartadas. Todos los análisis se realizaron en Python y están disponibles (Boté-Vericad, 2024).

En tercer lugar, se llevó a cabo un protocolo de descarga y análisis de datos de Factiva de la siguiente manera: los ficheros se descargaron en formato RTF, luego se convirtieron a formato TXT para eliminar caracteres no deseados o metadatos innecesarios. Seguidamente, se realizó una normalización de las noticias, tanto de la cabecera como del cuerpo del texto. Una vez que todo el corpus estuvo normalizado, se convirtió a formato CSV (valores separados por comas), lo que permitió estructurar la información y facilitar su análisis. Una vez obtenida la información estructurada, se procedió a realizar los siguientes análisis en Python. Se llevó a cabo un análisis de la distribución temporal por año y mes, así como un análisis autorías por periódico, y tendencias en el tiempo. Se obtuvieron frecuencias de palabras mediante nubes de palabras tanto de los títulos de las noticias como de los contenidos, y se generó un mapa de calor con las palabras más frecuentes a lo largo del tiempo. Además, se realizó un análisis de menciones de entidades nombradas (Vállez et al., 2024). Antes del análisis de datos, se llevaron a cabo varios procesos de depuración para asegurar su calidad y coherencia (Boté-Vericad, 2023).

Como parte de la minería de textos aplicada al corpus de noticias, se realizaron tres tipos de análisis utilizando técnicas de procesamiento de lenguaje natural. Primero, se llevó a cabo un análisis de sentimientos, categorizando los valores en positivo, negativo y neutro (Pang y Lee, 2008; Liu, 2012; Cambria et al., 2013). A continuación, se efectuó un análisis de correlación entre los periódicos y los resultados del análisis de sentimientos. El segundo análisis consistió en una exploración de los temas presentes en el corpus completo. Para ello, primero se calculó la coherencia de temas, una métrica que mide la consistencia semántica de los temas generados por el modelo. Esta métrica evalúa si las palabras agrupadas dentro de un tema específico coexisten frecuentemente en los documentos, lo que sugiere que el tema es coherente y representativo de una idea específica (Röder et al., 2015; Afolabi y Uzor, 2022; Aletras y Stevenson, 2013). En Python, esta métrica se puede calcular utilizando bibliotecas como *gensim*. Al graficar la coherencia de los temas en función del número de temas seleccionados, se busca identificar un ‹codo› o punto de inflexión, que generalmente indica el número óptimo de temas (Figura 8). Una vez determinado el número óptimo de temas, se realizó un análisis Latente de Dirichlet (LDA, por sus siglas en inglés) mediante un modelo preentrenado de aprendizaje automático para la exploración de dichos temas. El tercer análisis se enfocó en una red de co-ocurrencia de palabras para explorar los términos relacionados, utilizando el programa VOSviewer (Van Eck y Waltman, 2023) para dicho análisis.

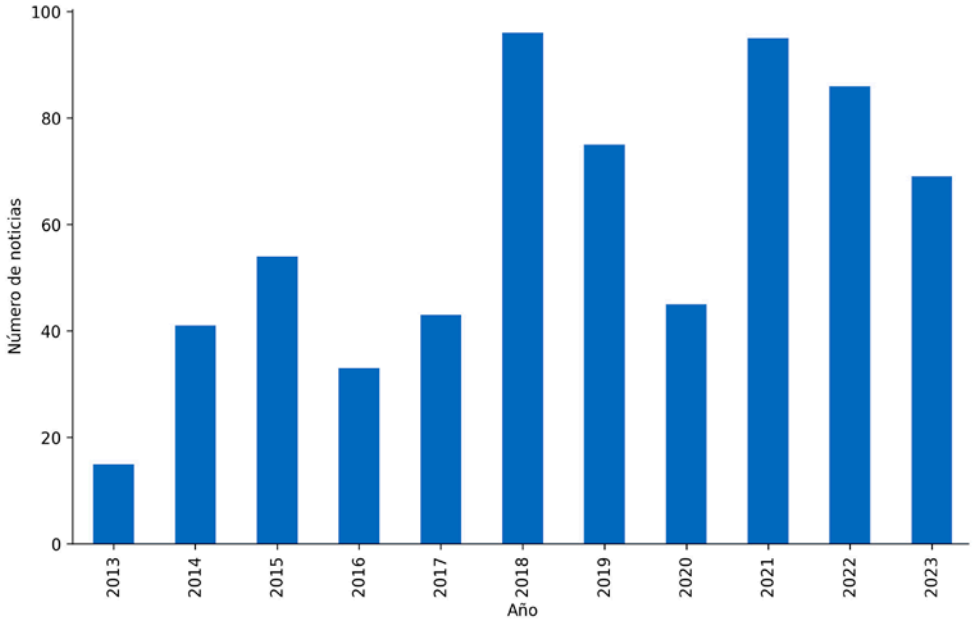
4. Resultados

A continuación, se presentan los resultados obtenidos mediante los métodos descritos en la metodología.

4.1. Distribución temporal de artículos por año

La Figura 1 presenta la evolución del número de noticias publicadas anualmente desde 2013 hasta 2023. Al observar la tendencia, se puede discernir un patrón de crecimiento inicial seguido de fluctuaciones significativas en los años más recientes. En 2013, el número de noticias es relativamente bajo, comenzando con menos de 20. Sin embargo, a partir de ese año, se observa un incremento constante en la cantidad de noticias publicadas. Este crecimiento se mantiene durante los años siguientes, alcanzando un punto culminante en 2018, cuando el número de noticias llega a 96, lo que marca el año con mayor actividad informativa dentro del período analizado. En 2019, hay una disminución en la cantidad de noticias, aunque esta caída no es tan pronunciada. Sin embargo, en 2020, la tendencia descendente se acentúa notablemente, y como consecuencia del COVID-19, podría sugerir un cambio en las dinámicas de producción o publicación de noticias en ese año específico.

Figura 1. Número de artículos por año



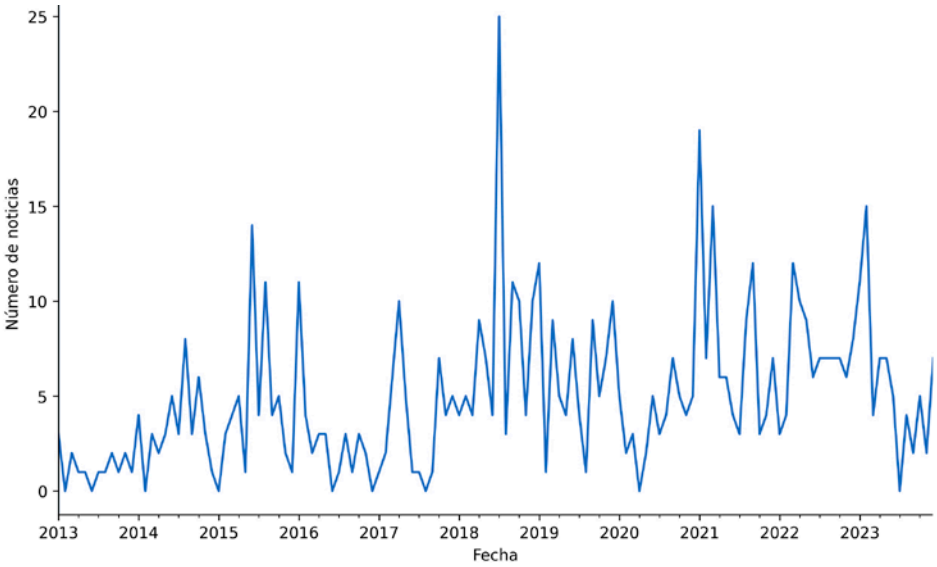
Fuente: Elaboración propia.

Contrariamente a esta caída, 2021 marca una recuperación significativa. El número de noticias vuelve a elevarse a 95, alcanzando niveles comparables a los de 2018, lo que indica una reactivación en la actividad informativa. Este aumento, sin embargo, no se mantiene de forma sostenida, ya que en 2022 se registra una ligera disminución con 86 noticias siendo considerablemente alto en comparación con los años anteriores. Finalmente, en 2023, se observa una nueva reducción en la cantidad de noticias hasta 69, sugiriendo una cierta estabilización tras las fluctuaciones de los años previos.

4.2. Tendencia temporal de noticias

Se ha realizado un análisis de la tendencia temporal de noticias (Figura 2). El eje de abcisas (Eje X) representa el tiempo, abarcando desde principios de 2013 hasta 2023. Cada punto en el eje de abcisas corresponde a un mes dentro de este rango temporal. El de ordenadas (Eje Y) muestra el número de noticias publicadas en cada mes. Los valores varían desde 0 hasta 25, indicando la cantidad de artículos publicados por mes. La línea de tendencia muestra cómo ha variado el número de artículos publicados cada mes a lo largo del tiempo.

Figura 2. Tendencia temporal de noticias sobre Wikipedia



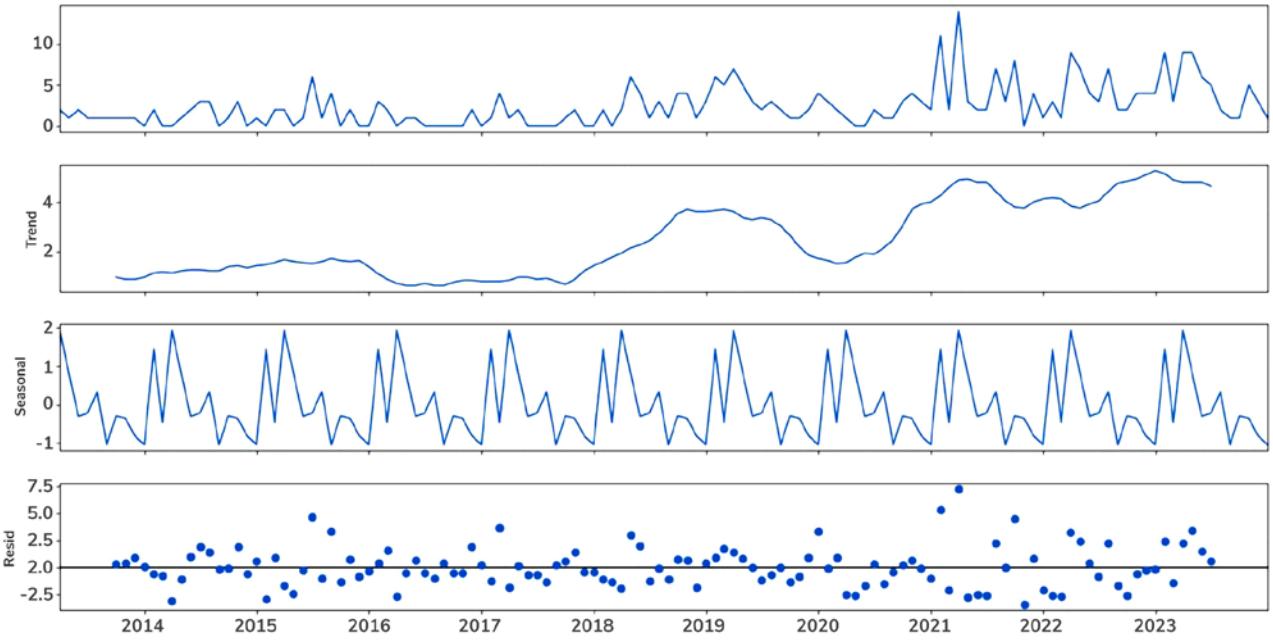
Fuente: Elaboración propia.

La gráfica de la Figura 2, muestra una serie de picos y valles a lo largo del tiempo, indicando fluctuaciones en el número de noticias publicadas. Hay un aumento inicial (2013-2015) en el que la cantidad de noticias publicadas parece aumentar de manera moderada y luego experimenta un primer pico significativo alrededor de 2015. Se observa una serie de fluctuaciones (2016-2018) con varios picos y valles. Hay un pico notable a finales de 2018. Se observa un aumento sostenido (2019-2021) que culmina en varios picos importantes durante 2020 y 2021, seguido de una caída. La cantidad de noticias publicadas parece estabilizarse (2022-2023) con fluctuaciones menores y picos menos pronunciados. La tendencia respecto de los picos de publicación sugiere que en 2015 estuvo asociado a eventos internacionales, cambios significativos en la política, o crisis económicas que generan un aumento en la cobertura mediática. Por ejemplo, el bloqueo de Wikipedia en Rusia, la demanda de Wikimedia contra la Agencia de Seguridad Nacional de EE.UU. (NSA) o la concesión del Premio Princesa de Asturias a Wikipedia.

En 2018 la crisis relacionada con la Directiva Europea sobre Derechos de Autor y la oposición global a la regulación europea puede haber impulsado la publicación de noticias (Quintais, 2019). Los picos entre 2020 y 2021 estarían relacionados con la pandemia de COVID-19 como evento clave que ha dominado las noticias globales durante estos años, resultando en un aumento significativo en la publicación de noticias relacionadas con la pandemia, políticas de salud, vacunas, y respuestas gubernamentales. Entre 2022 y 2023, en el contexto del conflicto en Ucrania, surge una crisis de credibilidad en Wikipedia, donde se descubrieron artículos ficticios sobre la historia de Rusia (Bradshaw et al., 2024; Szostek, 2018), creados a lo largo de una década, además de intentos de desinformación por parte de editores prorrusos. También se produjo el bloqueo de Wikipedia en Turquía por negarse a eliminar artículos relacionados con la guerra civil siria.

La gráfica de la Figura 3 corresponde a la descomposición estacional de la serie temporal de las noticias. La descomposición estacional desglosa la serie temporal en tres componentes: tendencia *(Trend)*, estacionalidad *(Seasonal)* y residuo *(Residual)*. La Serie Original *(Observed)* muestra el número de noticias publicadas cada mes donde los picos y valles en esta gráfica representan los momentos en que hubo un mayor o menor número de publicaciones de noticias. La Tendencia *(Trend)* muestra el patrón general a largo plazo en la serie temporal, suavizando las fluctuaciones a corto plazo para revelar el patrón general a largo plazo. La Tendencia muestra un aumento gradual en el número de noticias publicadas desde 2014 hasta alrededor de 2021, seguido de una leve disminución. Esto sugiere un incremento sostenido en la actividad de publicaciones de noticias durante este período, con una pequeña corrección posterior.

Figura 3. Descomposición estacional de la serie temporal de las noticias



Fuente: Elaboración propia.

La componente Estacionalidad *(Seasonal)* muestra patrones repetitivos en la serie temporal que ocurren a intervalos regulares, como los ciclos anuales, mensuales o semanales. La estacionalidad muestra patrones recurrentes, con picos y valles que se repiten cada año. Esto indica que hay ciertos meses o períodos del año en los que la publicación de noticias es consistentemente mayor o menor. Esto puede indicar que hay ciertos meses en los que se publican consistentemente más o menos noticias. La componente Residuo *(Residual)* representa las fluctuaciones aleatorias que quedan después de eliminar la tendencia y la estacionalidad de la serie original. Los residuos parecen estar distribuidos alrededor de cero, con algunos picos y valles. La dispersión de los puntos indica la presencia de variaciones aleatorias o anomalías que no siguen un patrón claro.

La tendencia muestra un aumento sostenido en el número de noticias publicadas desde 2014 hasta 2021. Esto podría estar relacionado con un mayor interés en ciertos temas, una mayor actividad en la plataforma, o eventos específicos que impulsaron la publicación de noticias. Los picos y valles regulares en el componente estacional indican la presencia de estacionalidad. Esto sugiere que ciertos períodos del año tienen consistentemente más o menos noticias publicadas.

A nivel práctico y en el contexto de la Wikipedia el aumento de la tendencia hasta 2021 indica un periodo de alta actividad en Wikipedia, lo que refleja la importancia creciente de la plataforma en la difusión de información. Esto puede ser aprovechado para fomentar la participación incentivando a más colaboradores a unirse, destacando la relevancia y el impacto de sus contribuciones. También a mejorar la infraestructura invirtiendo en mejores herramientas y recursos para soportar el aumento de actividad y asegurar que los servidores y la plataforma en general puedan manejar el tráfico y la carga de trabajo. La ligera disminución y estabilización post-2021 puede ser un indicativo de saturación de contenido.

4.3. Distribución de la frecuencia de palabras

El análisis de la frecuencia de palabras en las noticias publicadas en Wikipedia ofrece una visión integral de cómo los temas de interés han evolucionado a lo largo del tiempo. La Figura 4 muestra una nube de palabras que visualiza las frecuencias de palabras clave extraídas de los títulos de noticias relacionadas con Wikipedia. En la nube, las palabras más grandes, como «Wikipedia», «artículo», «página», «información» y «editor», son aquellas que aparecen con mayor frecuencia

en los títulos. Estas palabras resaltan los temas más recurrentes en la cobertura mediática, que incluyen la propia plataforma, la creación y edición de artículos, así como aspectos relacionados con la información y los editores. Otros términos relevantes, como «guerra», «Rusia», «Ucrania» y «bloqueo», sugieren la presencia de temas geopolíticos y controversias, mientras que términos como «mujer» y «derecho» apuntan a cuestiones de género y derechos en la plataforma. La visualización ofrece una visión clara de los temas predominantes en las noticias sobre Wikipedia durante el periodo analizado.

Figura 4. Frecuencia de palabras por títulos

Fuente: Elaboración propia.

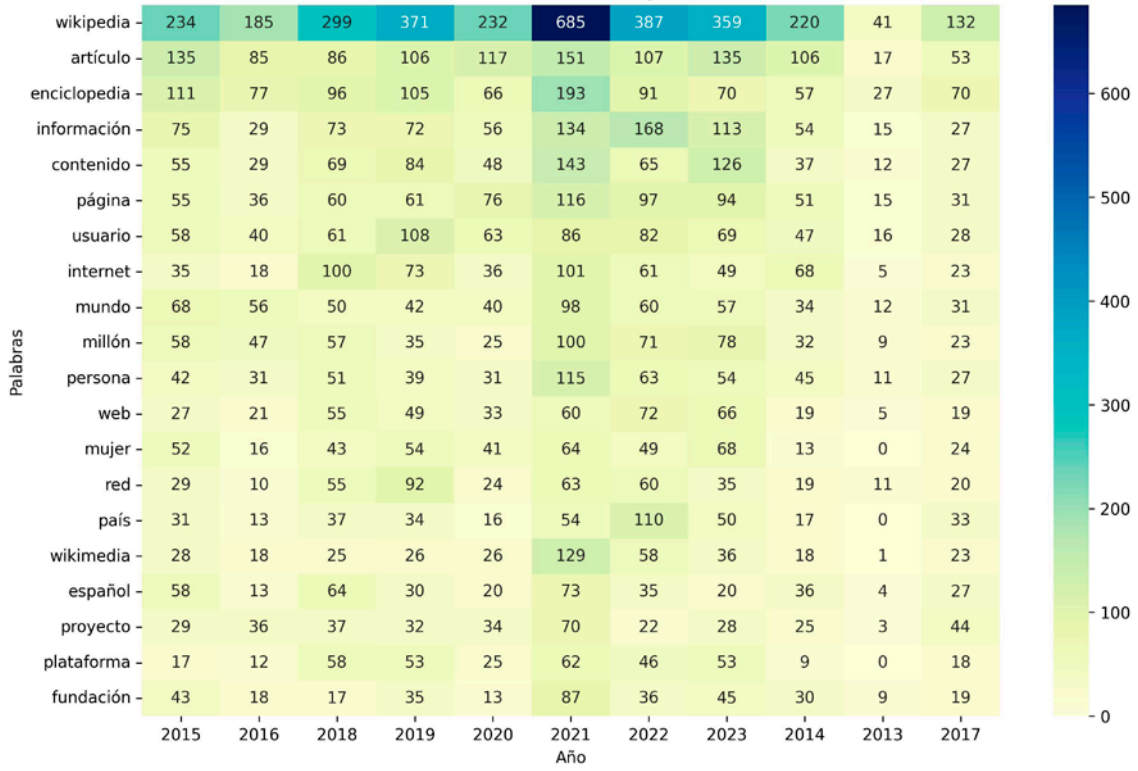
La Figura 5 presenta una nube de palabras basada en los contenidos de las noticias sobre Wikipedia. Las palabras más prominentes, como «Wikipedia», «artículo», «enciclopedia», «información» y «contenido», indican que estos términos son los más frecuentes en los textos. Otros términos destacados incluyen «internet», «mundo», «persona» y «editor», lo que sugiere un enfoque en la dimensión global de Wikipedia y la participación de individuos en la creación y edición de contenido. Esta visualización refleja las principales temáticas tratadas en los contenidos sobre Wikipedia, con énfasis en su función como fuente de información y plataforma colaborativa.

Figura 5. Frecuencia de palabras por contenidos

Fuente: Elaboración propia

Utilizando un mapa de calor (Figura 6), podemos visualizar las fluctuaciones en la aparición de términos clave, lo que permite identificar patrones y tendencias relevantes en el contenido de la plataforma. Este análisis se centra en la frecuencia de términos desde 2013 hasta 2023, proporcionando una perspectiva longitudinal sobre la dinámica de publicación de noticias en Wikipedia.

Figura 6. Mapa de calor de palabras



Fuente: Elaboración propia.

El mapa de calor presentado en este estudio muestra la selección de palabras más comunes y se calculó la frecuencia de aparición anual de cada término en el periodo de estudio (2013-2023). La intensidad del color en el mapa de calor refleja la frecuencia, con tonos más oscuros indicando una mayor frecuencia de aparición. Este enfoque visual facilita la identificación de picos y valles en el uso de términos específicos a lo largo del tiempo.

4.4. Palabras de Alta Frecuencia

«Wikipedia»: La palabra «Wikipedia» muestra una alta frecuencia constante a lo largo de todos los años, con un pico significativo en 2021. Este incremento notable puede estar asociado con un aumento en la atención mediática y la relevancia de la plataforma durante eventos globales importantes, como la pandemia de COVID-19.

«Artículo»: La frecuencia de la palabra «artículo» aumenta gradualmente, alcanzando su punto máximo en 2021. Esto sugiere un crecimiento en la creación y discusión de artículos en Wikipedia, posiblemente impulsado por la necesidad de información detallada y actualizada durante la pandemia.

«Enciclopedia»: La palabra «enciclopedia» presenta un patrón similar, con un pico en 2021. Este término refuerza la percepción de Wikipedia como una fuente de conocimiento enciclopédico, especialmente relevante en tiempos de crisis cuando se busca información confiable.

4.5. Palabras de Frecuencia Moderada

"Información": La palabra "información" tiene picos notables en 2021, reflejando la alta demanda de información precisa durante la pandemia de COVID-19. La frecuencia moderada pero constante de este término destaca la función de Wikipedia como un repositorio de información.

"Usuario": Este término muestra una alta frecuencia constante, indicando la importancia continua de la comunidad de Wikipedia en la creación y mantenimiento del contenido. Los picos pueden corresponder a campañas de contribución o a momentos de alta participación de la comunidad.

4.6. Palabras de Baja Frecuencia

«Mujer» y «Fundación»: Estos términos presentan frecuencias menores y más variables, sugiriendo que, aunque son temas de interés, su relevancia fluctúa más en función de eventos específicos o iniciativas particulares dentro de la plataforma.

«Internet» y «Mundo»: Las palabras relacionadas con tecnología y globalización muestran variabilidad en su frecuencia, reflejando cómo estos temas ganan y pierden relevancia en diferentes períodos.

Los resultados del mapa de calor revelan varias tendencias clave en la frecuencia de palabras en las noticias de Wikipedia:

Impacto de Eventos Globales: El pico en 2021 para términos como «Wikipedia», «artículo» e «información» está claramente asociado con la pandemia de COVID-19, que llevó a un aumento masivo en la búsqueda de información confiable y actualizada.

Crecimiento y Evolución: El aumento gradual en la frecuencia de términos como «artículo» y «enciclopedia» sugiere un crecimiento sostenido en la cantidad de contenido y en la percepción de Wikipedia como una fuente de conocimiento enciclopédico.

Participación Comunitaria: La consistencia en la frecuencia de la palabra «usuario» subraya la importancia de la comunidad de editores y colaboradores en Wikipedia, y los picos pueden reflejar momentos de alta actividad comunitaria.

Temas Variables: La variabilidad en la frecuencia de términos como «mujer», «fundación», «internet» y «mundo» indica que ciertos temas tienen una relevancia más episodica, dependiendo de eventos o campañas específicas.

El mapa de calor de la frecuencia de palabras en las noticias de Wikipedia permite visualizar y entender las tendencias en la plataforma a lo largo del tiempo. Los picos en la frecuencia de términos clave pueden ser directamente correlacionados con eventos globales significativos, mientras que la consistencia en otros términos refleja la evolución y el crecimiento continuo de Wikipedia. Este análisis no solo destaca la dinámica de la publicación de noticias en Wikipedia, sino que también ofrece percepciones valiosas para guiar futuras estrategias editoriales y de contenido, asegurando que la plataforma siga siendo una fuente relevante y confiable de información para usuarios de todo el mundo.

4.7. Análisis de entidades nombradas

En la Tabla 2 se presenta un ranking de las 20 principales entidades mencionadas. Destaca la predominancia de entidades tecnológicas y geopolíticas en el contexto de la cobertura mediática sobre Wikipedia. Wikipedia (625 menciones) y Wikimedia (167 menciones) son las entidades más referenciadas, lo que subraya la centralidad de la enciclopedia digital y su fundación en el discurso de los medios de habla hispana. Compañías tecnológicas como Google o Facebook, y

Microsoft en menor medida, también aparecen con frecuencia, lo que refuerza el vínculo entre el acceso a la información y las plataformas tecnológicas. Asimismo, las menciones a países como Rusia, Ucrania, Francia, y EE.UU. sugieren un fuerte enfoque en el contexto geopolítico, especialmente en relación con conflictos y el manejo de la información.

En cuanto a las figuras personales, Jimmy Wales y Larry Sanger, cofundadores de Wikipedia, continúan destacándose como protagonistas. Entre las personalidades políticas, sobresalen Vladimir Putin y Donald Trump, ambos mencionados en relación con debates sobre la información y la desinformación. También destacar la autoridad de supervisión rusa de los medios de comunicación Roskomnadzor (13 menciones) en el conflicto de Ucrania y la eliminación de contenido en Wikipedia. Un aspecto interesante es la presencia de Patricia Horrillo (7 menciones), activista de Wikipedia y fundadora de Wikiesfera, quien aparece junto a Florencia Claes, presidenta de Wikimedia España (7 menciones). Esto refleja su implicación en el fomento de la edición en Wikipedia desde un enfoque comunitario. No obstante, la representación femenina sigue siendo limitada, con Cristina Kirchner (7 menciones) como la primera mujer política en el ranking, ocupando la posición 25, lo que sugiere una escasa presencia femenina en el discurso mediático sobre Wikipedia.

Tabla 2. Menciones de Entidades en la Cobertura Mediática sobre Wikipedia en la Prensa Hispanoablante

Ranking	Apariciones	Entidad en el Texto	Entidad
1	625	Wikipedia	ORG
2	167	Wikimedia	ORG
3	116	Google	ORG
4	94	Facebook	ORG
5	60	Jimmy Wales	PER
6	57	Madrid	LOC
7	54	Chile	LOC
8	47	Larry Sanger	PER
9	46	Europa	LOC
10	43	Rusia	LOC
11	38	Ucrania	LOC
12	33	Francia	LOC
13	14	Vladimir Putin	PER
14	14	EE.UU	LOC
15	14	Donald Trump	PER
16	13	Roskomnadzor	ORG
17	11	Elon Musk	PER
18	10	Wikitribune	ORG
19	7	Patricia Horrillo	PER
20	7	Florencia Claes	PER
	...	.....	
25	7	Cristina Kirchner	PER

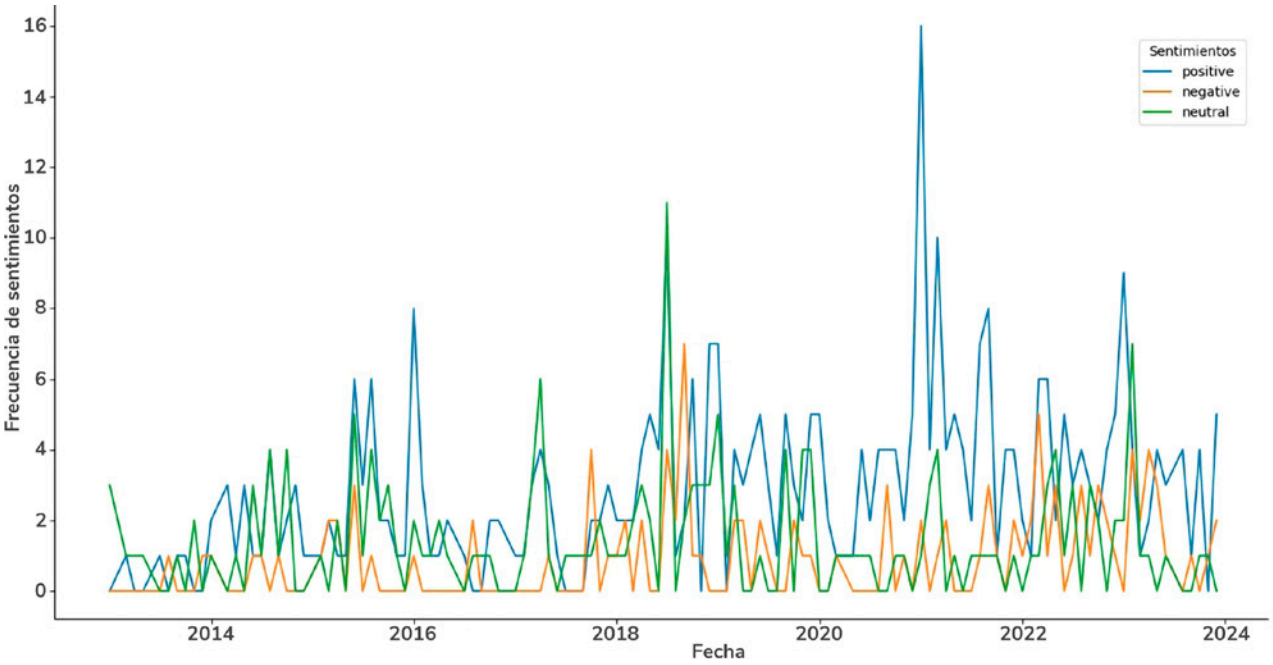
Fuente: Elaboración propia. ORG: Organizaciones, PER: Personas, LOC: Localizaciones.

4.8. Análisis de sentimientos

Se ha realizado un análisis de sentimientos a lo largo del tiempo empleando la librería '*Textblob*' de Python. Dado el volumen de datos se han empleado sólo tres estados positivo, negativo y neutro (Liu, 2012; Pang y Lee, 2008). Se puede observar que la Figura 7 muestra la evolución

temporal de la frecuencia de sentimientos positivos, negativos y neutrales en las noticias desde 2013 hasta 2023. A través de esta visualización, podemos identificar patrones y tendencias en la tonalidad de las noticias a lo largo del tiempo.

Figura 7. Evolución del análisis de sentimientos en el tiempo



Fuente: Elaboración propia.

Predominio de Sentimientos Positivos: En el período comprendido entre 2020 y 2022, se observa un claro predominio de noticias con sentimientos positivos. Estos picos sugieren que durante estos años, las noticias fueron más optimistas o se enfocaron en aspectos positivos, posiblemente como una respuesta a ciertos eventos o cambios en la cobertura mediática.

Fluctuaciones de Sentimientos Negativos: Aunque las noticias negativas parecen ser menos frecuentes que las positivas en general, hay momentos en los que se observan picos significativos de negatividad, particularmente en 2018 y a principios de 2022. Estos picos pueden estar asociados con eventos específicos que generaron una percepción negativa en el discurso mediático, como crisis, conflictos o noticias de alto impacto negativo.

Presencia de Sentimientos Neutrales: A lo largo del período analizado, también se identifican noticias con sentimientos neutros. Sin embargo, la frecuencia de estos parece ser más baja y dispersa en comparación con los sentimientos positivos y negativos. Esto podría indicar que, aunque existen noticias neutrales, el enfoque de la cobertura tiende a inclinarse hacia posiciones más emocionales, ya sea en un sentido positivo o negativo.

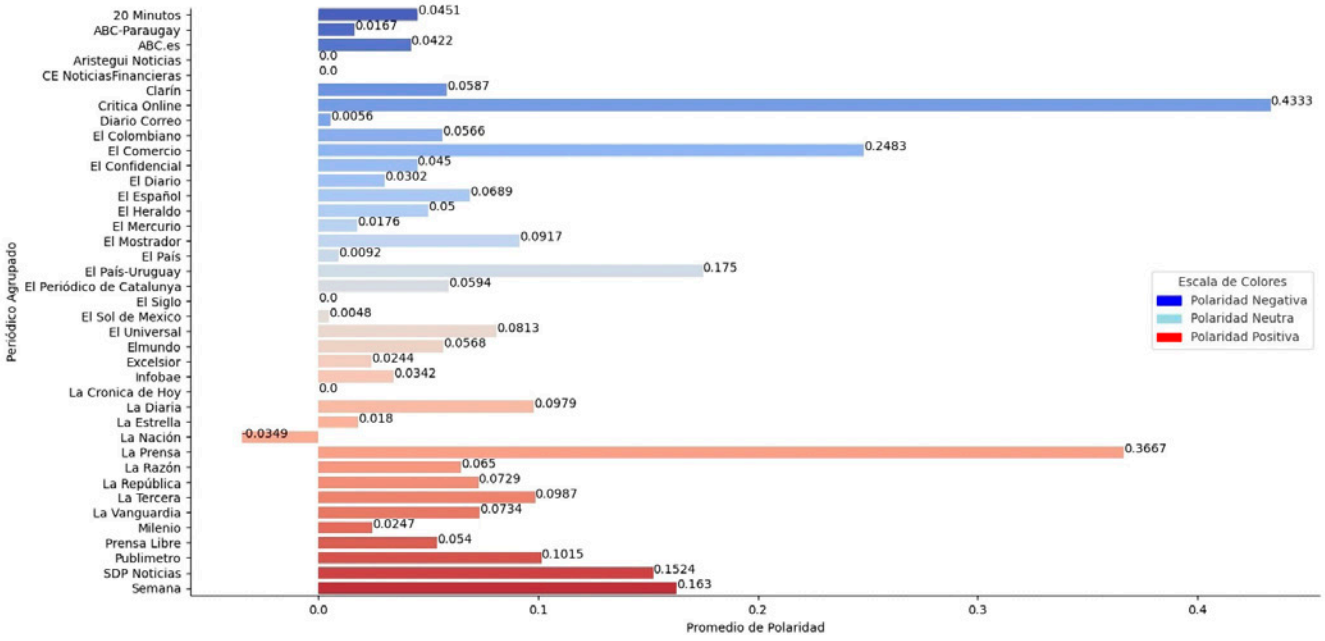
Los altos picos de sentimientos negativos en 2018 y principios de 2022 sugieren que en esos años se produjeron eventos que llevaron a una cobertura más negativa. Estos picos pueden reflejar momentos de crisis, controversias o situaciones de gran repercusión mediática. Ejemplos son el COVID-19 y el conflicto bélico de Ucrania que está presente en nuestra muestra estudiada.

El predominio de sentimientos positivos durante 2021 hasta mitad de 2022 podría estar relacionado con la cobertura de temas que generaron esperanza, recuperación o avances positivos, especialmente considerando el contexto de la pandemia de COVID-19, donde ciertas noticias sobre vacunas, recuperación económica, o solidaridad podrían haber sido predominantes.

4.9. Análisis de correlación entre periódicos digitales y análisis de sentimientos

En la Figura 8 se muestra la correlación entre los periódicos digitales y los sentimientos expresados en las noticias. El total de periódicos digitales analizados es de 41, con un promedio de 16 noticias por medio y una mediana de 6 noticias. Esto indica que existe una distribución asimétrica, probablemente con una larga cola hacia la derecha, lo que sugiere que algunos periódicos publicaron muchas más noticias que otros. Aun así, el promedio de 16 noticias por periódico sugiere que, en general, los periódicos publican un volumen de noticias relativamente alto. La mediana de 6 noticias por periódico, significativamente menor que la media, refuerza la idea de una distribución asimétrica. Es por esto que se ha realizado una correlación de Spearman así como un análisis de regresión robusta.

Figura 8. Correlación de los periódicos digitales en función de los sentimientos



Fuente: Elaboración propia.

En el análisis de correlación de Spearman realizado para evaluar la relación entre la polaridad media de los artículos y el número de artículos publicados por cada medio de comunicación agrupado, se obtuvo un coeficiente de correlación de Spearman de 0,0171 con un valor p de 0,918 (Tabla 2). Estos resultados indican que no existe una correlación significativa entre la polaridad media de los artículos y el número de artículos publicados. El valor de correlación cercano a cero sugiere que las variaciones en la polaridad media no están relacionadas de manera monótona con el número de artículos publicados.

Se llevó a cabo un análisis de regresión robusta como se puede observar en la Figura 9, para explorar la relación entre la polaridad media de los artículos y el número de artículos publicados por los medios de comunicación agrupados. Los resultados, presentados en la Tabla 3, mostraron un coeficiente de -27,28 para la variable de polaridad media, con un valor p de 0,182. Aunque el coeficiente negativo sugiere una relación inversa, el valor p indica que esta relación no es estadísticamente significativa a un nivel convencional ($p < 0,05$). Esto implica que no hay evidencia suficiente para concluir que la polaridad de los artículos influye de manera significativa en la cantidad de artículos publicados. Estos hallazgos sugieren que la variabilidad en el tono o sentimiento de los artículos no afecta sustancialmente el volumen de noticias producidas por los medios. Esto podría indicar que los factores que determinan la cantidad de contenido publicado son independientes del tono emocional del mismo, o que la polaridad del contenido no es un criterio decisivo en las decisiones editoriales de los medios analizados.

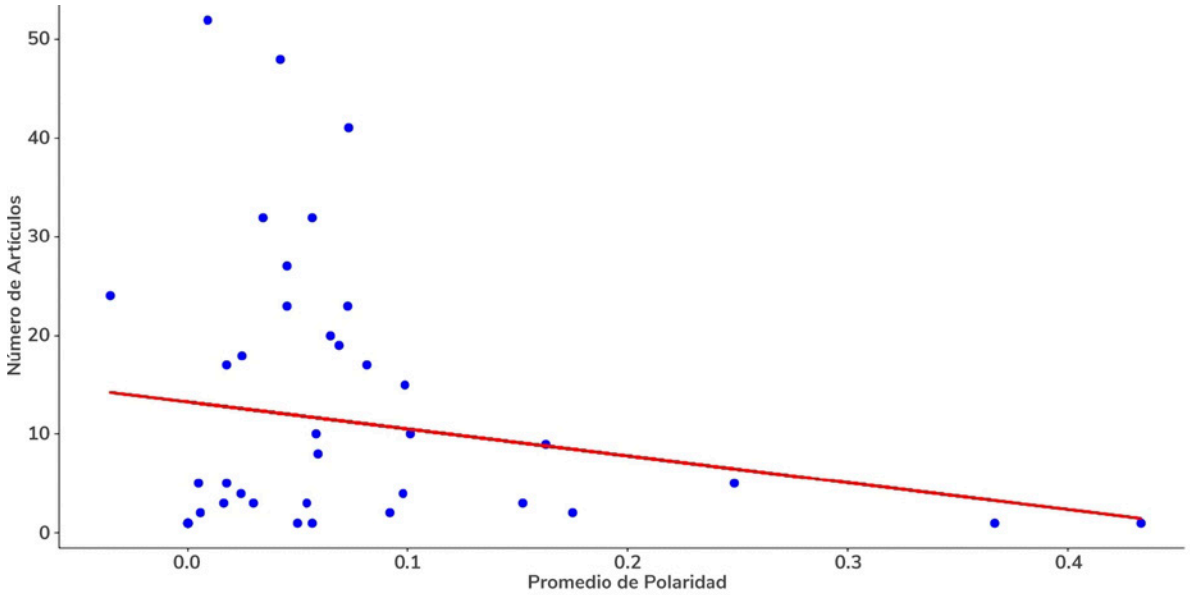
El modelo de regresión robusta que explora la relación entre la polaridad media de los artículos y el número de artículos publicados por los medios de comunicación agrupados se puede expresar matemáticamente como:

$$\text{Número de artículos} = 13{,}2156 - 27{,}2772 \times \text{Polaridad Media} + \epsilon$$

Donde el coeficiente de la polaridad media es -27,2772 (valor p = 0,182), lo que sugiere una relación inversa, aunque no estadísticamente significativa, entre la polaridad y el número de artículos publicados. El término de error ϵ captura las variaciones no explicadas por el modelo.

Finalmente, los resultados de los dos análisis indican que el tono o sentimiento general de los artículos, tal como se mide en este estudio, no parece tener un impacto discernible en la cantidad de contenido producido por estos medios

Figura 9. Análisis de regresión robusta



Fuente: Elaboración propia.

Tabla 3. Tabla de Resultados de la Regresión Robusta y correlación de Spearman

Variable	Coeficiente	Error estándar	Z-Valor	Valor p	Intervalo de confianza 95%
Constante	13,2156	2,446	5,402	0,000	[8,421, 18,011]
Polaridad Media	-27,2772	20,438	-1,335	0,182	[-67,335, 12,781]
Correlación de Spearman: 0,0171055, p-value: 0,9176804					

Fuente: Elaboración propia.

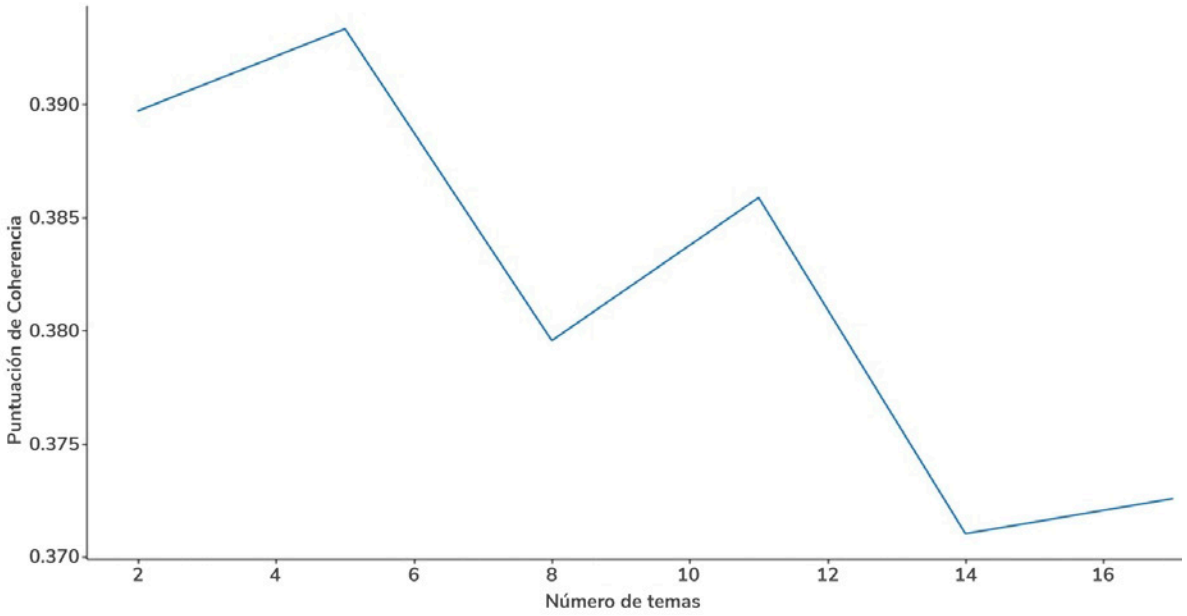
4.10. Análisis temático

Como se ha mencionado previamente se ha calculado la métrica de coherencia de temas mediante la librería "gensim" de Python. Como se puede observar en la figura 10, el número óptimo resultante es de 5.

La Figura 11 presenta una visualización de las tendencias temáticas que han emergido a lo largo del tiempo en relación con Wikipedia. Este análisis ha sido realizado utilizando un modelo preentrenado para identificar y categorizar los temas principales del corpus. Los temas identificados reflejan diversas dimensiones del desarrollo, funcionamiento, y percepción de Wikipedia en el periodo estudiado.

Desarrollo y Crecimiento de Wikipedia (Línea Azul): Este tema abarca la evolución de Wikipedia desde sus inicios hasta su estado actual. Las fluctuaciones en la frecuencia de este tema a lo largo del tiempo podrían estar relacionadas con hitos importantes en la historia de la plataforma, como la introducción de nuevas características, mejoras en la infraestructura tecnológica, o expansiones en la cobertura de contenidos. Picos específicos en la línea azul sugieren momentos clave donde el crecimiento de la plataforma fue particularmente destacado en el corpus.

Figura 10. Número óptimo de temas



Fuente: Elaboración propia.

Proyecto y Alcance Global de Wikipedia (Línea Naranja): Este tema se centra en cómo Wikipedia ha expandido su alcance a nivel global, incluyendo la creación de versiones en múltiples idiomas y su impacto en diferentes regiones del mundo. La variabilidad en la frecuencia de este tema refleja el interés en cómo Wikipedia ha logrado convertirse en una herramienta esencial en la diseminación del conocimiento a nivel mundial. Los aumentos en esta tendencia pueden correlacionarse con campañas globales de la plataforma o eventos que resaltan su impacto internacional.

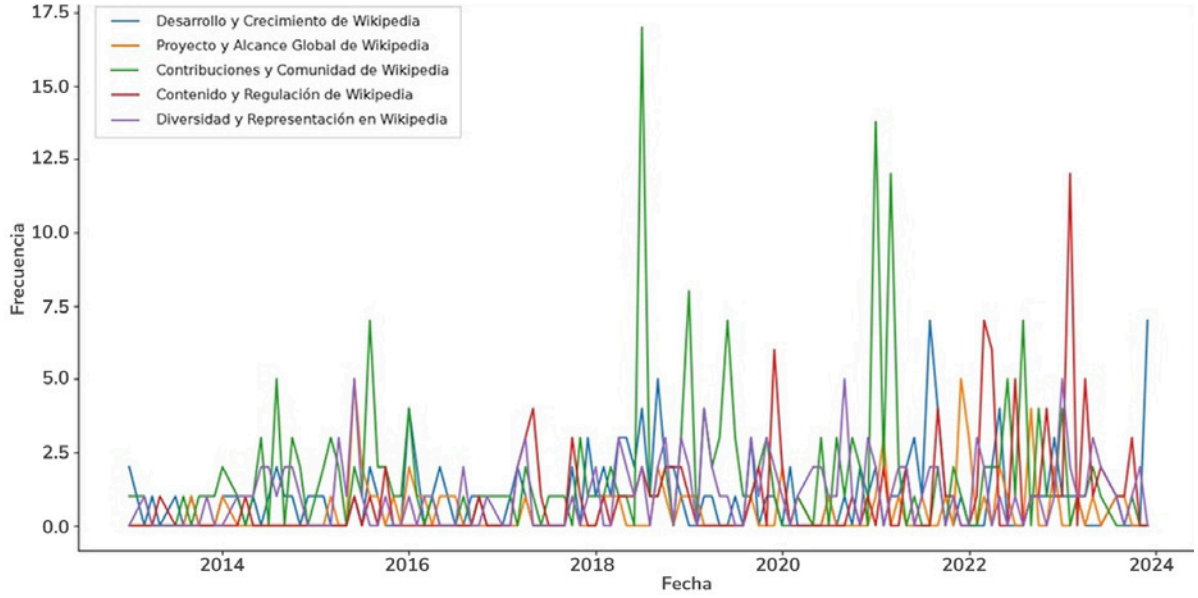
Contribuciones y Comunidad de Wikipedia (Línea Verde): Este tema destaca la importancia de la comunidad de editores y colaboradores que sustentan el funcionamiento de Wikipedia. Los picos significativos en la línea verde pueden estar asociados con eventos que fomentaron la participación de los usuarios, como editatonas, debates sobre la gobernanza comunitaria, o crisis que requerían una respuesta coordinada por parte de la comunidad. La variabilidad en este tema subraya la naturaleza colaborativa y la dinámica de participación dentro de Wikipedia.

Contenido y Regulación de Wikipedia (Línea Roja): Este tema aborda las políticas y procedimientos que regulan la creación, edición y mantenimiento del contenido en Wikipedia. La aparición de picos en esta línea puede estar vinculada a discusiones o controversias sobre la precisión del contenido, decisiones editoriales importantes, o cambios en las políticas de Wikipedia que impactan la manera en que se regula y controla la información disponible en la plataforma. Esto refleja las continuas negociaciones y ajustes necesarios para mantener la fiabilidad y neutralidad de Wikipedia.

Diversidad y Representación en Wikipedia (Línea Morada): Finalmente, el tema de la diversidad y representación en Wikipedia se centra en cómo diferentes grupos y perspectivas están representados en la plataforma. La presencia de picos en esta línea puede estar asociada con

debates sobre la representación de géneros, culturas, y otras identidades en los artículos de Wikipedia. Este tema es particularmente relevante en el contexto de esfuerzos globales para mejorar la inclusión y representación justa en el conocimiento accesible a través de Wikipedia.

Figura 11. Tendencias temáticas en el corpus analizado

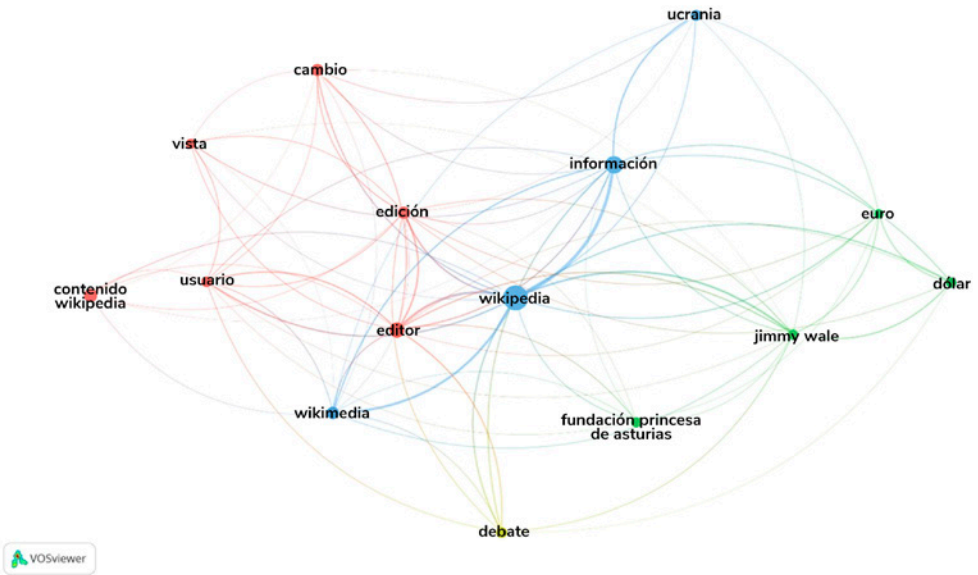


Fuente: Elaboración propia.

4.11. Red de co-ocurrencia de palabras por contenidos

La Figura 12 presenta una red de co-ocurrencias de palabras generada a partir de un corpus relacionado con Wikipedia. Este tipo de análisis es fundamental para identificar y comprender las conexiones temáticas dentro de un conjunto de datos textuales. La red se compone de nodos, que representan palabras, y aristas, que indican la frecuencia con la que estos términos co-ocurren en el mismo contexto. Además, los nodos están organizados en agrupaciones o clústeres diferenciados por colores, los cuales reflejan las principales áreas temáticas emergentes del corpus.

Figura 12. Análisis de co-ocurrencias por contenidos



Fuente: Elaboración propia.

El análisis de la red revela cuatro agrupaciones temáticas bien definidas, cada una destacando diferentes aspectos del uso y la gestión de la información en Wikipedia.

En primer lugar, el clúster representado en color verde está centrado en la figura de Jimmy Wales, uno de los cofundadores de Wikipedia. Este clúster conecta a Wales con la Fundación Princesa de Asturias, institución que le otorgó un reconocimiento en 2015, subrayando su relevancia en el ámbito de la cultura y la información global. Dentro del mismo clúster se encuentran los términos dólar y euro, que están asociados con las solicitudes de donaciones que la plataforma realiza regularmente a sus usuarios. Esta asociación sugiere que las discusiones en torno a Wales también incluyen aspectos financieros cruciales para la sostenibilidad de Wikipedia. En segundo lugar, el clúster azul posiciona a Wikipedia como un nodo central, destacando su papel crucial como fuente de información. Este clúster está estrechamente relacionado con temas de relevancia internacional, como el conflicto bélico en Ucrani". La'prominencia de Ucrania en este clúster sugiere que Wikipedia ha sido una fuente importante de información durante este evento geopolítico, reflejando su papel en la diseminación de conocimientos en tiempos de crisis global.

En tercer lugar, el clúster rojo está intrínsecamente ligado a la edición de contenidos en Wikipedia. Este grupo de palabras resalta la actividad de los usuarios como editores, una función esencial en la estructura colaborativa de la plataforma. Los términos "cambio" y "edición" refuerzan la idea de que Wikipedia es una plataforma dinámica, donde el contenido es constantemente revisado y actualizado por su comunidad de usuarios. Este clúster subraya la importancia de la participación activa de los usuarios en la creación y mantenimiento de la información. Finalmente, el clúster amarillo, aunque más reducido, se conecta con los tres clústeres anteriores y está centrado en el término "debate". En Wikipedia, el proceso de debate es fundamental para garantizar la calidad y neutralidad de los artículos antes de su aprobación final. Este clúster refleja cómo las discusiones y consensos dentro de la comunidad son cruciales para la validez y fiabilidad de la información presentada en la plataforma.

5. Discusión

Los resultados del análisis de la cobertura mediática de Wikipedia en la prensa hispanohablante entre 2013 y 2023 ofrecen varias perspectivas que contrastan con la revisión de la literatura. En primer lugar, la evolución temporal de las noticias sobre Wikipedia muestra fluctuaciones significativas, con picos en años específicos como 2018 y 2021, lo que puede estar relacionado con eventos globales como las elecciones políticas y la pandemia de COVID-19. Esto coincide con lo observado por Lee (2018), quien encontró que la dinámica del establecimiento de la agenda entre medios de comunicación y Wikipedia refleja los intereses y preocupaciones del momento. Los resultados obtenidos pueden interpretarse a través de la Teoría del Encuadre, que sugiere que la manera en que los medios presentan la información afecta directamente la percepción del público. Por ejemplo, el aumento de la cobertura positiva durante 2021 podría estar relacionado con un encuadre que resaltaba la importancia de Wikipedia como fuente confiable durante la pandemia de COVID-19. En contraste, los picos de negatividad en 2018 podrían indicar un encuadre que destacaba controversias o desafíos asociados a la plataforma. Esto demuestra que los medios no solo informan sobre Wikipedia, sino que también moldean la narrativa en torno a la enciclopedia, lo que puede influir en la percepción pública y en la forma en que Wikipedia es utilizada como recurso.

En cuanto a la PI1, los resultados del análisis de la frecuencia de palabras y el mapa de calor revelan que los términos «Wikipedia», «artículo» e «información» son algunos de los más recurrentes, especialmente en momentos clave como 2021. Este patrón refleja la consolidación de Wikipedia como una fuente de conocimiento crucial durante eventos globales, como la pandemia de COVID-19. La prominencia de estos términos también sugiere que los medios han enmarcado Wikipedia como un recurso confiable y relevante en tiempos de crisis, lo

que coincide con la percepción de Wikipedia como una herramienta indispensable para la diseminación de información, tal como lo destacaron Piñeiro-Naval y Mangana (2018). Al respecto de la PI2, el análisis de sentimientos realizado muestra un predominio de noticias con sentimientos positivos entre 2020 y 2022, lo que podría indicar que durante estos años, los medios hispanohablantes enmarcaron las noticias sobre Wikipedia de manera más optimista, posiblemente para reflejar la utilidad de la plataforma en un contexto de incertidumbre global. Los picos de sentimientos negativos en 2018 y 2022, por otro lado, pueden estar asociados con la cobertura de controversias o conflictos, sugiriendo que los medios también enmarcaron a Wikipedia de manera crítica en ciertos contextos. Este enfoque de encuadre, que destaca lo positivo o negativo según el contexto, es consistente con la Teoría del Encuadre, que sostiene que los medios estructuran la información para influir en la percepción pública, como lo discutió Pérez-Salazar (2019).

Finalmente, en la PI3, el análisis de la red de co-ocurrencias de palabras revela la existencia de varios clústeres temáticos, como el relacionado con Jimmy Wales y la Fundación Princesa de Asturias, o el que destaca la relevancia de Wikipedia en el contexto del conflicto en Ucrania. Estos clústeres indican que los medios no solo cubren a Wikipedia de manera superficial, sino que también la relacionan con figuras clave, eventos globales y temas de gran relevancia. Este enfoque sugiere que los medios enmarcan estas noticias para destacar la importancia de Wikipedia en contextos específicos, lo que se alinea con la Teoría del Encuadre, que explica cómo los medios seleccionan y enfatizan ciertos aspectos de una noticia para influir en la percepción del público (Zheng, 2020).

6. Limitaciones

Una de las limitaciones de este estudio es la dependencia exclusiva de la base de datos Factiva, lo que podría excluir noticias relevantes de medios no incluidos en esta base. Además, el análisis se centra únicamente en artículos que mencionan a Wikipedia en el título, lo que podría haber dejado fuera contenido relevante incluido en el cuerpo de los textos que no cumple con este criterio específico. La aplicación de técnicas de procesamiento de lenguaje natural y *machine learning* también presenta desafíos, como la clasificación imprecisa de sentimientos, la agrupación temática y la necesidad de validación adicional para asegurar la precisión de los modelos utilizados. Aunque la Teoría del Encuadre aporta una perspectiva valiosa, futuros estudios podrían complementarla con análisis de agenda mediática para explorar las interacciones entre Wikipedia y los medios. Finalmente, sería útil expandir el análisis a otras lenguas y contextos culturales para una comprensión más amplia de las dinámicas mediáticas en torno a Wikipedia.

7. Conclusiones

Este estudio evidencia cómo la cobertura mediática de Wikipedia en la prensa hispanohablante entre 2013 y 2023 ha respondido a eventos globales específicos. Durante la pandemia de COVID-19, se registró un aumento significativo de noticias positivas, enmarcando a Wikipedia como una fuente confiable en momentos de incertidumbre. Del mismo modo, el conflicto en Ucrania destacó su papel en la difusión de información durante crisis geopolíticas. En contraste, picos de sentimientos negativos en 2018 estuvieron vinculados con controversias como el debate sobre la Directiva Europea de Derechos de Autor. El análisis temático y de sentimientos revela una predominancia de noticias positivas durante eventos clave, reflejando el valor de Wikipedia como repositorio de información colaborativa. Sin embargo, las noticias negativas, aunque menos frecuentes, se asocian con controversias sobre representatividad y fiabilidad, como los sesgos de género. También se identifica la influencia de la prensa española y su papel en destacar debates editoriales y narrativas geopolíticas. Los hallazgos subrayan la capacidad de los medios para influir en la percepción pública mediante encuadres específicos. Además, plantean la necesidad de fortalecer iniciativas para mitigar sesgos y fomentar la diversidad en los contenidos de Wikipedia, promoviendo su relevancia como herramienta colaborativa y confiable.

8. Contribuciones

El autor confirma que es el único responsable de todos los procesos de la investigación: conceptualización, análisis formal, administración del proyecto, investigación, metodología, tratamiento de datos, recursos, software, supervisión, validación, visualización de resultados, redacción.

Bibliografía

Afolabi, I. T., y Uzor, C. N. (2022). Topic Modelling for Research Perception: Techniques, Processes and a Case Study. En M. Al-Emran y K. Shaalan (Eds.), *Recent Innovations in Artificial Intelligence and Smart Applications* (pp. 221-237). Springer. https://doi.org/10.1007/978-3-031-14748-7_13

Aletras, N., y Stevenson, M. (2013). Evaluating Topic Coherence Using Distributional Semantics. En A. Koller y K. Erk (Eds.), *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (pp. 13-22). Association for Computational Linguistics. https://aclanthology.org/W13-0102

Boté-Vericad, J.-J. (2024). *Códigos Python Análisis de Contenidos de Noticias sobre Wikipedia en la prensa hispanoablante*. Zenodo. https://doi.org/10.5281/zenodo.13827464

Boté-Vericad, J.-J. (2023). *Integrating mixed methods to analyse information behaviour in the use of educational videos in higher education* [Stiftung Universität Hildesheim]. https://doi.org/10.25528/141

Boté-Vericad, J.-J. (2022). *Analysis of Spotify Spanish spoken profiles in Twitter*. https://doi.org/10.5281/zenodo.6618902

Bradshaw, S., Elswah, M., Haque, M., y Quelle, D. (2024). Strategic storytelling: Russian state-backed media coverage of the Ukraine war. *International Journal of Public Opinion Research*, *36*(3), edae028. https://doi.org/10.1093/ijpor/edae028

Cambria, E., Schuller, B., Xia, Y., y Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. *IEEE Intelligent Systems*, *28*(2), 15-21. IEEE Intelligent Systems. https://doi.org/10.1109/MIS.2013.30

Carmel, E. (2013). Mobility, migration and rights in the European Union: critical reflections on policy and practice. *Policy Studies*, *34*(2), 238–253. https://doi.org/10.1080/01442872.2013.778028

Casado-Gutiérrez, F., Sánchez, R., Luque González, A., y García Guerrero, J. E. (2021). La pandemia Covid-19 según los medios internacionales: El caso de Ecuador a través de la teoría del framing en Twitter. *RISTI: Revista Ibérica de Sistemas e Tecnologias de Informação*, *Extra 40*, 410-422. https://www.risti.xyz/issues/ristie40.pdf#page=59

Chhabra, A., y Iyengar, S. R. S. (2020). Who Writes Wikipedia? An Investigation from the Perspective of Ortega and Newton Hypotheses. *Proceedings of the 16th International Symposium on Open Collaboration*, 1-11. https://doi.org/10.1145/3412569.3412578

Debus, M., y Florczak, C. (2022). Using party press releases and Wikipedia page view data to analyse developments and determinants of parties' issue prevalence: Evidence for the right-wing populist 'Alternative for Germany'. *Research & Politics*, *9*(3). https://doi.org/10.1177/20531680221116570

Ferran-Ferrer, N., Boté-Vericad, J.-J., y Minguillón, J. (2023). Wikipedia gender gap: A scoping review. *El Profesional de la información*, e320617. https://doi.org/10.3145/epi.2023.nov.17

Gluza, W., Turaj, I., y Meier, F. (2021). Wikipedia Edit-a-thons and Editor Experience: Lessons from a Participatory Observation. *Proceedings of the 17th International Symposium on Open Collaboration*, 1-9. https://doi.org/10.1145/3479986.3479994

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harper & Row.

Hinnosaar, M. (2019). Gender inequality in new media: Evidence from Wikipedia. *Journal of Economic Behavior & Organization*, *163*, 262-276. https://doi.org/10.1016/j.jebo.2019.04.020

Hobolt, S. B., Leeper, T. J., y Tilley, J. (2021). Divided by the vote: Affective polarization in the wake of the Brexit referendum. *British Journal of Political Science*, *51*(4), 1476–1493. https://doi.org/10.1017/S0007123420000125

Johnson, G., Anderson, C., Dunning, K., y Williamson, R. (2024). National ocean policy in the United States: Using framing theory to highlight policy priorities between presidential administrations. *Frontiers in Marine Science*, *11*. https://doi.org/10.3389/fmars.2024.1370004

Kaffee, L.-A., Arora, A., y Augenstein, I. (2023). Why should this article be deleted? Transparent stance detection in multilingual Wikipedia editor discussions. En H. Bouamor, J. Pino, y K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5891–5909). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.361

Keswani, K., Das, I., Shrivastava, B., Gupta, A., y Katarya, R. (2020). LDA based model for mining textual features from financial news articles. En *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 43–48). https://doi.org/10.1109/ICACCCN51052.2020.9362882

Krishnamoorthy, S. (2018). Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems*, *56*(2), 373-394. https://doi.org/10.1007/s10115-017-1134-1

Lee, P. T. Y. (2018). In search of public agenda with text mining: An exploratory study of agenda setting dynamics between the traditional media and Wikipedia. En M. Ganji, L. Rashidi, B. C. M. Fung, y C. Wang (Eds.), *Trends and applications in knowledge discovery and data mining* (pp. 309–317). Springer. https://doi.org/10.1007/978-3-030-04503-6_30

Liu, B. (2012). *Sentiment analysis and opinion mining*. Springer. https://doi.org/10.1007/978-3-031-02145-9

Liu, C. (2020). Analysis of Relationship Between Hot News and Stock Market——Based on LDA Model and Event Study. *Journal of Physics: Conference Series*, *1616*(1), 012048. https://doi.org/10.1088/1742-6596/1616/1/012048

Messner, M., y South, J. (2011). LEGITIMIZING WIKIPEDIA: How US national newspapers frame and use the online encyclopedia in their coverage. *Journalism Practice*, *5*(2), 145-160. https://doi.org/10.1080/17512786.2010.506060

Mishra, A., Sahay, A., Pandey, M. A., y Routaray, S. S. (2023). News text analysis using text summarization and sentiment analysis based on nlp. *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*, 28-31. https://doi.org/10.1109/ICSMDI57622.2023.00014

Morris-O'Connor, D., Strotmann, A., y Zhao, D. (2022). Editorial Behaviors for Biasing Wikipedia Articles. *Proceedings of the Association for Information Science and Technology*, *59*(1), 226-234. https://doi.org/10.1002/pra2.618

Muñiz, C. (2011). Encuadres noticiosos sobre migración en la prensa digital mexicana: Un análisis de contenido exploratorio desde la teoría del framing. *Convergencia*, *18*(55), 213-239.

Mutua, S. N., y Oloo, D. (2020). Online news media framing of COVID-19 pandemic: Probing the initial phases of the disease outbreak in international media. *European Journal of Interactive Multimedia and Education, 1*(2), e02006. https://doi.org/10.2139/ssrn.4667716

Pang, B., y Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval. *Foundations and Trends in Information Retrieval*, *21*(2), 1-135.

Pérez-Salazar, G. (2019). Teoría del encuadre y plataformas sociodigitales de interacción: Un análisis de coyuntura. *Revista Mexicana de Ciencias Políticas y Sociales, 64*(236), 333–353. https://doi.org/10.22201/fcpys.2448492xe.2019.236.68820

Petroni, F., Broscheit, S., Piktus, A., Lewis, P., Izacard, G., Hosseini, L., Dwivedi-Yu, J., Lomeli, M., Schick, T., Bevilacqua, M., Mazaré, P.-E., Joulin, A., Grave, E., y Riedel, S. (2023). Improving Wikipedia verifiability with AI. *Nature Machine Intelligence*, *5*(10), 1142-1148. https://doi.org/10.1038/s42256-023-00726-1

Pinto, R., Lacerda, J., Silva, L., Araújo, A. C., Fontes, R., Lima, T. S., Miranda, A. E., Sanjuán, L., Gonçalo Oliveira, H., Atun, R., y Valentim, R. (2023). Text mining analysis to understand the impact of online news on public health response: Case of syphilis epidemic in Brazil. *Frontiers in Public Health*, *11*, 1248121. https://doi.org/10.3389/fpubh.2023.1248121

Piñeiro-Naval, V., y Mangana, R. (2018). Teoría del encuadre: Panorámica conceptual y estado del arte en el contexto hispano. *Estudios sobre el Mensaje Periodístico*, *24*(2), 1541–1557. https://doi.org/10.5209/ESMP.62233

Piñeiro-Naval, V.; Igartua, J.-J., Marañón-Lazcano, F. de J., y Sánchez-Nuevo, A. (2018). El análisis de contenido y su aplicación a entornos web: un caso empírico. Tendencias metodológicas en la investigación académica sobre comunicación. *Espejo De Monografías De Comunicación Social*, (2), 253–272. https://doi.org/10.52495/c6.2.emcs.2.mic6

Prasad, O. J., Nandi, S., Dogra, V., y Diwakar, D. S. (2023). A systematic review of NLP methods for Sentiment classification of Online News Articles. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 1-9. https://doi.org/10.1109/ICCCNT56998.2023.10308056

Ptaszek, G., Yuskiv, B., y Khomych, S. (2024). War on frames: Text mining of conflict in Russian and Ukrainian news agency coverage on Telegram during the Russian invasion of Ukraine in 2022. *Media, War & Conflict*, *17*(1), 41-61. https://doi.org/10.1177/17506352231166327

Quintais, J. P. (2019). The new copyright in the digital single market directive: A critical look. *European Intellectual Property Review, 2020*(1). https://doi.org/10.2139/ssrn.3424770

Ren, R., y Xu, J. (2024). It's not an encyclopedia, it's a market of agendas: Decentralized agenda networks between Wikipedia and global news media from 2015 to 2020. *New Media & Society, 26*(11), 6235-6259. https://doi.org/10.1177/14614448221149641

Röder, M., Both, A., y Hinneburg, A. (2015). Exploring the space of topic coherence measures. En *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). https://doi.org/10.1145/2684822.2685324

Sádaba, T. (2001). Origen, aplicación y límites de la "teoría del encuadre" (framing) en comunicación. *Comunicación y Sociedad*, *14*, 143-175. https://doi.org/10.15581/003.14.36373

Shao, D., Li, C., Huang, C., Xiang, Y., y Yu, Z. (2022). A news classification applied with new text representation based on the improved LDA. *Multimedia Tools and Applications, 81*(15), 21521–21545. https://doi.org/10.1007/s11042-022-12713-6

Silva, L., y Barbosa, L. (2022). Matching news articles and Wikipedia tables for news augmentation. *Knowledge and Information Systems, 65*(4), 1713–1734. https://doi.org/10.1007/s10115-022-01815-0

Sv, S. B., y Geetha, A. (2019). Determination of news biasedness using content sentiment analysis algorithm. *Indonesian Journal of Electrical Engineering and Computer Science, 16*(2), 882–889. https://doi.org/10.11591/ijeecs.v16.i2.pp882-889

Szostek, J. (2018). Nothing is true? The credibility of news and conflicting narratives during "information war" in Ukraine. *The International Journal of Press/Politics, 23*(1), 116–135. https://doi.org/10.1177/1940161217743258

Urologin, S. (2018). Sentiment analysis, visualization and classification of summarized news articles: A novel approach. *International Journal of Advanced Computer Science and Applications, 9*(8), 616–625. https://doi.org/10.14569/IJACSA.2018.090878

Valera-Ordaz, L. (2016). El sesgo mediocéntrico del 'framing' en España: Una revisión crítica de la aplicación de la teoría del encuadre en los estudios de comunicación. *ZER: Revista de Estudios de Comunicación, 21*(40), 13–30. https://doi.org/10.1387/zer.17259

Vállez, M., Boté-Vericad, J.-J., Guallar, J., y Bastos, M. T. (2024). Indifferent about online traffic: The posting strategies of five news outlets during musk's acquisition of twitter. *Journalism Studies*, *25*(11), 1249-1271. https://doi.org/10.1080/1461670X.2024.2372437

Van Eck, N. J., y Waltman, L. (2023). *VOSviewer* (Version 1.6.20) [Computer software]. https://www.vosviewer.com

Walter, S. (2019). Better off without you? How the British media portrayed EU citizens in Brexit news. *The International Journal of Press/Politics, 24*(2), 210–232. https://doi.org/10.1177/1940161218821509

Wirawan, R., Krisnanik, E., y Arista, A. (2024). Text mining for news forecasting on the Turnback Hoax website. *JOIV: International Journal on Informatics Visualization, 8*(1), 96–106. https://doi.org/10.62527/joiv.8.1.1939

Yang, P., y Colavizza, G. (2024). Polarization and reliability of news sources in Wikipedia. *Online Information Review, 48*(5), 908–925. https://doi.org/10.1108/OIR-02-2023-0084

Yang, Y., Kaizhong, J., Mingjun, Y., & Laxin, H. (2022). Selecting optimal LDA numbers to identify news topics. *Data Analysis and Knowledge Discovery, 6*(11), 72–78. https://doi.org/10.11925/infotech.2096-3467.2022.0115

Yin, R. K. (2003). Case Study Methodology. En *Case Study Research Design and Methods* (3.ª ed., pp. 96–106). Sage.

Zheng, S. (2020). The communication power of Chinese novel coronavirus pneumonia (COVID-19) news reports in light of the framing theory. *Theory and Practice in Language Studies, 10*(11), 1467–1473. https://doi.org/10.17507/tpls.1011.18