
Special issue

Javier Cantón-Correa

<https://orcid.org/0000-0002-8466-1679>

javier.canton@unir.net

Universidad Internacional de La Rioja

Mariola Moreno

<https://orcid.org/0000-0001-5624-6905>

mmorenoc@unav.es /

mariola.moreno@unir.net

Universidad Internacional de La Rioja /

Instituto de Cultura y Sociedad,

Universidad de Navarra

Recibido

3 de diciembre de 2025

Aprobado

11 de mayo de 2026

© 2026

Communication & Society

ISSN 0214-0039

E ISSN 2386-7876

www.communication-society.com

2026 – Vol. 19 (1)

pp. 482-500

Cómo citar este artículo:

Cantón-Correa, J., & Moreno, M. (2026).

Verificación colaborativa en X: Del fact-checking profesional a las Notas de la Comunidad, *Communication & Society*, 39(1), 482-500

<https://doi.org/10.15581/003.39.1.032>

Verificación colaborativa en X: del fact-checking profesional a las Notas de la Comunidad

Resumen

La transición hacia modelos colaborativos está transformando la gestión de la verificación de desinformación en redes sociales. Este estudio analiza todas las Notas de la Comunidad en español publicadas en X, examinando su evolución temporal, distribución temática, patrones de participación, tiempos de aprobación y funcionamiento del consenso mediante análisis descriptivo y modelado de tópicos. Los resultados muestran un crecimiento sostenido del sistema y una concentración temática en política, economía, seguridad y migración. La participación sigue una distribución de tipo ley de potencia, con un núcleo reducido de usuarios muy activos y una mayoría amplia de colaboradores ocasionales. Los tiempos de aprobación presentan demoras significativas derivadas del algoritmo de consenso transversal, lo que limita la intervención temprana. El sistema opera como un mecanismo complementario de detección y contextualización, aunque su eficacia se ve condicionada por la polarización temática, la desigualdad participativa y la lentitud del consenso.

Palabras clave

Notas de la Comunidad, fact-checking, desinformación, redes sociales, verificación colaborativa, gobernanza algorítmica.

1. Introducción

En el ecosistema mediático actual, la circulación paulatina de contenidos no verificados en las redes sociales constituye un desafío crucial tanto para las sociedades democráticas como para las instituciones mediáticas. Este problema se agrava en un contexto de creciente desconfianza hacia las instituciones sociales, incluidos los medios tradicionales, los verificadores de información y las plataformas digitales. En España, el último informe publicado por Digital News Report 2025 (Reuters Institute, 2025) indica que la confianza en los medios tradicionales ha caído al nivel más bajo de la última década, sólo un 31% de los encuestados afirmó confiar en ellos. Esta fragmentación en la confianza mediática e institucional contribuye a la pérdida de la capacidad de articular narrativas compartidas y verificables. A ello se suma que los algoritmos de distribución de contenido priorizan la rapidez, el volumen y la carga emocional por encima de criterios epistemológicos o de verificación, lo que facilita la rápida propagación de bulos y afirmaciones engañosas (Cantón-Correa, 2024; Wardle & Derakhshan, 2017).

Así pues, desde la creación de FactCheck.org en 2003 en Estados Unidos, la verificación de datos ha experimentado una expansión significativa en el contexto internacional, consolidándose como una estrategia clave para mitigar los desórdenes informativos (Kyriakidou

et al., 2022). Este crecimiento se refleja en el registro del Duke Reporters' Lab, que contabiliza 443 proyectos activos de verificación en todo el mundo en 2025. Estos verificadores profesionales surgieron como una respuesta institucional clave frente a la desinformación masiva, ya que analizan declaraciones públicas, comprueban datos y detectan errores utilizando metodologías sistemáticas, con transparencia y colaboración internacional (Montoro-Montarroso et al., 2023). En el ámbito hispanohablante, esta labor ya está consolidada, tanto en España como en Latinoamérica existen organizaciones reconocidas, como Maldita.es, Newtral, Verificat y EFE Verifica en España, o Chequeado en Argentina, que han adaptado las metodologías de verificación al contexto local, colaborado con redes internacionales y creando repositorios de datos de desinformación en español.

No obstante, la literatura señala límites estructurales de este modelo de verificación profesional en el ecosistema digital actual, apuntando, por un lado, a la limitada capacidad que tienen para cubrir el volumen masivo de contenido generado por los usuarios, ya que los métodos tradicionales no pueden seguir el ritmo de difusión de la desinformación, puesto que los rumores y bulos se propagan más rápido que las correcciones (Allen, et al., 2021), y por otro lado, la brecha temporal entre la aparición de un contenido falso y su posterior verificación reduce la eficacia de esta última, dado que las correcciones suelen llegar cuando el material ya ha alcanzado una amplia circulación (Chuai et al., 2024).

A este desfase entre la información veraz y verificada y los desórdenes informativos, se suma la pérdida de la confianza pública y la percepción de sesgo partidista, que afectan negativamente la recepción de las verificaciones, así como la dependencia financiera de alianzas con plataformas digitales, cuyas políticas pueden cambiar de manera abrupta y comprometer la sostenibilidad del sector (Westlund et al., 2024). Estas tensiones han impulsado la búsqueda de modelos alternativos o complementarios que permitan aumentar la escala y rapidez de la corrección informativa, incluyendo herramientas basadas en inteligencia artificial para la detección temprana de contenidos potencialmente desinformativos (Cantón-Correa et al., 2025b), así como modelos de moderación comunitaria y verificación colaborativa impulsada por usuarios.

Así pues, las plataformas digitales, convertidas en los principales canales de información pública, han adoptado en los últimos años políticas de moderación más laxas, externalizando parte del control de contenidos y experimentando con modelos de supervisión más descentralizados. EL 7 de enero de 2025, Meta anunció el reemplazo de los verificadores de datos en Estados Unidos por un sistema de notas de la comunidad, siguiendo un modelo que X ya había implementado cuatro años antes. Además, la compañía simplificó sus políticas de contenido y eliminó ciertas restricciones (Meta, 2025). Este giro provocó reacciones en otras empresas tecnológicas, Google, por ejemplo, redujo sus compromisos de verificación de datos ante la Unión Europea (Google, 2025), mientras que TikTok, condicionó la implementación de procesos de verificación a la participación de empresas externas que ofrecieran servicios equivalentes, en lugar de asumir directamente esta función (TikTok, 2025). Estas decisiones pueden afectar tanto a la consistencia como en el alcance de las estrategias frente a la desinformación.

Más allá de los cambios técnicos, esta transición refleja un desplazamiento en la gobernanza de la información, donde, usuarios, algoritmos y plataformas pasan a desempeñar un papel central en la mediación de la verdad pública, generando nuevas tensiones en torno a la responsabilidad, la transparencia y la legitimidad.

1.1. La verificación profesional

La verificación de datos se ha consolidado como un movimiento profesional y un estilo innovador de periodismo, cuyo propósito es fortalecer la credibilidad informativa al exigir rendición de cuentas a los actores públicos frente a la propagación de falsedades (Graves, 2016). Así, como destaca Singer (2021), los verificadores de datos conciben su labor como un medio para fortalecer la autonomía ciudadana y promover la alfabetización mediática. Además, Amazeen (2020) apunta que también puede ser una herramienta para fortalecer la democracia y responder a la percepción de fragilidad o amenaza de las instituciones democráticas, lo cual es de suma importancia, en la medida en que los ciudadanos dispongan de información fiable para la toma de decisiones, tanto en el ámbito público y electoral, como en su vida personal, especialmente en el ámbito de la salud (Cantón-Correa, 2024). Además, en los últimos años, tanto las grandes empresas tecnológicas como los responsables políticos han incrementado la financiación destinada a iniciativas e instituciones de verificación (European Commission, 2023).

Es igualmente relevante subrayar que, aunque periodistas y verificadores comparten el objetivo de garantizar rigor y veracidad, sus funciones difieren sustancialmente. Mientras los periodistas se centran en asegurar la precisión antes de la publicación, los verificadores intervienen sobre contenidos ya difundidos, evaluando su autenticidad. Esta labor, según Singer (2021), trasciende la noción tradicional de objetividad e implica una búsqueda explícita y sistemática de evidencias. Se observa también un proceso de hibridación en las competencias de los fact-checkers, reflejado en la conformación de equipos multidisciplinares que integran especialistas de diversas áreas, más allá del ámbito periodístico (Moreno-Gil & Salgado-de Dios, 2023). Del mismo modo, las organizaciones de verificación pueden constituirse como entidades independientes o mantenerse vinculadas a medios de comunicación.

En España, la labor de verificación comenzó a adquirir especial relevancia en contextos electorales, donde el afloramiento de la desinformación favoreció el desarrollo de estas iniciativas. En este sentido, la investigación realizada por Poch-Butler et al. (2024) señala la preponderancia de la política como eje temático en los contenidos verificados en España. Asimismo, situaciones de crisis como la pandemia de la COVID-19 marcaron un momento de especial intensidad para la actividad de las plataformas verificadoras.

1.1.1. Desafíos de la verificación profesional

A pesar de su papel en la promoción de la veracidad informativa, los verificadores de datos han sido objeto de críticas persistentes. Entre ellas, se incluyen el sesgo en la selección de contenidos y el posible partidismo, habitualmente asociado a posiciones liberales. Además, enfrentan desafíos operativos como la sobrecarga cognitiva y la presión temporal, factores que incrementan su vulnerabilidad a sesgos cognitivos (Masotina et al., 2024). A ello se suman la posible inconsistencia en sus evaluaciones, que puede generar confusión, y la tendencia a simplificar en exceso declaraciones políticas complejas y, con frecuencia, ambiguas. Por otra parte, diversas investigaciones señalan que el partidismo político puede limitar la eficacia de las correcciones, al dificultar la convergencia de opiniones sobre el contenido verificado y aumentar la probabilidad de rechazo de la información rectificadora (Walter & Tukachinsky, 2020).

Asimismo, para garantizar la transparencia y la exhaustividad del trabajo de los verificadores profesionales, el Poynter Institute creó en 2015 la International Fact-Checking Network (IFCN), que establece compromisos de transparencia sobre las fuentes, la financiación, la estructura organizativa y los procedimientos de verificación (Dafonte-Gómez et al., 2021).

Desde entonces, los verificadores han reforzado su cooperación para consolidar su legitimidad profesional mediante un código ético y un sistema de acreditación que, a mediados de 2025, suma 182 organizaciones certificadas en 57 países (Nyariki, 2025). Las entidades adheridas se someten a una evaluación anual que verifica el cumplimiento del Código de Principios de la IFCN en materia de imparcialidad, transparencia y rendición de cuentas. En España, destacan como parte de esta red, Maldita.es, Newtral, Verificat y EFE Verifica (Cantón-Correa et al., 2025a).

1.2. La verificación colaborativa

La verificación colaborativa ha adquirido relevancia como mecanismo complementario al modelo profesional. Se fundamenta en la noción de la “sabiduría de las multitudes”, que sostiene que grandes grupos de usuarios pueden, bajo ciertas condiciones, aportar juicios útiles sobre la fiabilidad de los contenidos (Allen et al., 2021). Estudios recientes indican que, aunque los usuarios no expertos pueden no ser precisos de manera individual, la agregación de sus valoraciones produce un rendimiento cercano al de expertos (Allen et al., 2021; Martel et al., 2024) y permite ampliar la supervisión informativa en términos de escala y velocidad. No obstante, su eficacia depende de que los participantes presenten diversidad ideológica, independencia en sus juicios, acceso a fuentes fiables y protección frente a intentos de manipulación coordinada. Algunas investigaciones advierten que, en ausencia de mecanismos de gobernanza adecuados, la verificación colaborativa puede reproducir sesgos, amplificar errores o ser capturada por intereses organizados (Drolsbach et al., 2024), lo que subraya la necesidad de modelos híbridos que integren comunidad, algoritmos y análisis profesional.

En este contexto, las Notas de la Comunidad constituyen un ejemplo paradigmático de verificación colaborativa integrada en una plataforma social. Este sistema se originó en Twitter, que a comienzos de 2021 implementó el *Birdwatch* como primera iniciativa a gran escala en este ámbito, y posteriormente lo renombró Notas de la Comunidad (Wojcik et al., 2022). El funcionamiento del sistema permite que los colaboradores propongan anotaciones para tuits que consideran engañosos o carentes de contexto, sometiendo dichas propuestas a un proceso de evaluación comunitaria y algorítmica (Allen et al., 2022). En un inicio, este proceso presentó limitaciones sustanciales, entre ellas, la vulnerabilidad frente a intentos de manipulación coordinada, así como la presencia de sesgos partidistas que afectaban tanto al estilo de redacción como a los procesos de aprobación de las notas, aunque estos aspectos han ido mejorando. Si bien este mecanismo representa un cambio de paradigma, la cuestión ya no se limita únicamente a qué se verifica, sino también a quién lo hace y cómo esas correcciones llegan al público.

Desde su implementación se han observado efectos medibles, como el tiempo transcurrido entre la creación de una nota y su publicación efectiva, así como la disminución significativa de las interacciones y la difusión del tuit asociado (Slaughter et al., 2025). A pesar de estos avances, persisten problemas relevantes, dado que muchas anotaciones propuestas nunca alcanzan visibilidad y, en numerosos casos, se publican demasiado tarde para frenar la propagación del contenido engañoso, lo que limita su capacidad para interrumpir la dinámica de viralización (Chuai et al., 2024). La plataforma X ha destinado recursos significativos a optimizar el sistema mediante el desarrollo de un algoritmo avanzado que evalúa la utilidad de las notas y prioriza aquellas respaldadas por un conjunto diverso de usuarios, en lugar de una mayoría simple. Asimismo, ha incorporado criterios de elegibilidad para los colaboradores con el fin de garantizar la participación de usuarios auténticos.

Así pues, la dinámica de Notas de la Comunidad se articula en torno a tres componentes principales: la producción abierta de anotaciones por parte de los usuarios, la valoración

comunitaria de la utilidad de dichas anotaciones y un algoritmo de consenso que prioriza aquellas respaldadas por usuarios provenientes de distintos segmentos ideológicos (Wojcik et al., 2022). Este diseño busca promover la diversidad de perspectivas y evitar que un único grupo partidista controle la visibilidad de las correcciones. Sin embargo, la literatura advierte que este tipo de gobernanza algorítmica plantea desafíos importantes para la moderación, ya que la exigencia de un “puente ideológico” puede retrasar la publicación de notas o incluso impedirla por completo, lo que genera dudas sobre la legitimidad del sistema, la persistencia de sesgos residuales y la eficacia real de la corrección colaborativa en entornos polarizados.

1.2.1. Desafíos de las Notas de la Comunidad

La polarización constituye una de las principales amenazas para la verificación colaborativa. El trabajo de Allen et al. (2022) indica que los usuarios tienden a evaluar con menor generosidad las notas elaboradas por personas que perciben como ideológicamente opuestas, lo que reduce la posibilidad de alcanzar consensos transversales. Esta dinámica se intensifica cuando el sistema exige diversidad ideológica, ya que, en contextos polarizados, muchas anotaciones permanecen invisibles aun cuando sean técnicamente correctas. En este escenario, la gobernanza algorítmica desempeña un papel central al definir qué se considera diversidad suficiente, consenso o utilidad, y en esas decisiones se concentra parte de la autoridad epistémica del sistema. Cuando los umbrales se establecen sin transparencia o sin adaptación al contenido polarizado, pueden producirse efectos adversos como demoras, invisibilidad de correcciones legítimas o privilegio de contenidos dominantes.

Si bien las Notas de la Comunidad han mostrado resultados alentadores en contextos donde la moderación con verificadores externos es limitada, como la cobertura de contenido (Drolsbach et al., 2024), su consideración como solución definitiva, se basa en supuestos excesivamente optimistas sobre la integridad, la diversidad y la eficiencia de la colaboración entre usuarios. Tal como ocurrió con la promesa inicial del periodismo ciudadano de democratizar la información, que posteriormente fue cuestionada por problemas de capacidad, fiabilidad y ausencia de estándares profesionales, limitaciones similares podrían manifestarse en las Notas de la Comunidad.

Este trabajo se articula en tres ejes teóricos: (1) la transformación del fact-checking profesional en un contexto de crisis informativa y mediática; (2) la emergencia de la verificación colaborativa como modelo complementario y distribuido; y (3) la influencia del diseño algorítmico, la participación comunitaria y la polarización en la efectividad de estos sistemas.

2. Objetivos y preguntas de investigación

Este estudio tiene como objetivo evaluar de forma rigurosa el funcionamiento del sistema de las Notas de la Comunidad en español, analizando dimensiones como la frecuencia y la evolución de las anotaciones, los tiempos de aprobación, los tipos de errores corregidos, los patrones de participación de los colaboradores (en términos de actividad, persistencia y especialización temática), y su interacción con el ecosistema profesional del fact-checking, representado por organizaciones como Maldita.es o Chequeado. Al centrarse en un contexto hispanohablante, prácticamente inexplorado, esta investigación contribuye a cerrar una brecha en la literatura internacional y ofrece elementos de diagnóstico que pueden orientar la práctica de plataformas, verificadores y reguladores en materia de moderación digital.

Para ello, se parte de las siguientes preguntas de investigación:

1. ¿Cuál es la dinámica temporal y la frecuencia de publicación de las notas en español dentro del sistema de Notas de la Comunidad?

2. ¿Qué tipos de errores informativos abordan dichas anotaciones y cuál es su distribución temática en el ámbito español?
3. ¿Cuáles son los patrones de participación de los colaboradores hispanohablantes (en términos de actividad, persistencia y especialización temática)?
4. ¿En qué medida los temas abordados por las notas en español coinciden con las áreas informativas que suelen cubrir los verificadores profesionales?
5. ¿Cómo opera el sistema de consenso (votos de utilidad, diversidad ideológica de valoradores, tiempos de aprobación) en el entorno hispanohablante?

3. Metodología

A partir de las preguntas de investigación se diseñó una estrategia metodológica basada en la explotación integral de los conjuntos de datos oficiales de Notas de la Comunidad publicados por X (Twitter) en su repositorio público. Para este estudio se analizaron todos los registros disponibles desde el inicio del programa en 2021 hasta comienzos de 2025, lo que permite cubrir la fase completa de desarrollo del sistema y observar su evolución temporal en el medio plazo.

Los archivos se publican en formato TSV y recogen prácticamente toda la actividad del sistema: creación de notas, valoraciones de otros usuarios, cambios de estado interno, proceso de incorporación de colaboradores y señales generadas sobre tuits candidatos a recibir una nota. En concreto, se utilizaron cinco ficheros principales:

- **notes.tsv**: contiene todas las notas publicadas por los participantes e incluye información sobre la autoría, la marca temporal, el resumen redactado, la clasificación, los indicadores de corrección factual y las categorías de error informativo.
- **ratings.tsv**: reúne todas las evaluaciones realizadas por los usuarios sobre una nota específica. Incluye la marca temporal, el identificador del participante evaluador, el nivel de utilidad asignado (HELPFUL, SOMEWHAT_HELPFUL o NOT_HELPFUL) y un conjunto de variables binarias que indican los motivos de la evaluación, como claridad, calidad de las fuentes, información contextual, irrelevancia, errores o presencia de discurso sesgado.
- **notesStatusHistory.tsv**: registra la evolución del estado de cada nota dentro del sistema, lo que permite conocer si se encuentra actualmente considerada útil, si necesita más evaluaciones o si ha sido calificada como no útil.
- **userEnrollment.tsv**: describe el proceso de inscripción y la progresión de los participantes en el programa Notas de la Comunidad, incluyendo los cambios de estado y el número de ocasiones en que han alcanzado el nivel de confianza requerido por el sistema.
- **batSignals.tsv**: recoge todas las señales emitidas sobre un tuit que podrían dar lugar a la creación de una nota, junto con la información sobre quién las genera y el momento en que se producen.

A partir de estos ficheros se diseñó un flujo de trabajo orientado a integrar la información y hacerla manejable para el análisis. En una primera fase, los datos originales se almacenaron en la nube (AWS S3) y se consultaron mediante una base de datos analítica (AWS Athena), lo que permitió trabajar directamente sobre los archivos TSV sin necesidad de descargarlos ni transformarlos manualmente. Sobre esta base se generaron tablas derivadas en formato Parquet, más eficientes para el análisis, en las que se vincularon las distintas fuentes mediante identificadores comunes: el `noteId` para conectar notas y valoraciones, y el `tweetId` para relacionar notas y señales asociadas al mismo tuit.

Con estas tablas se construyeron varias estructuras analíticas que condensan la información relevante para el estudio: una tabla de notas enriquecida (con sus evaluaciones agregadas y su estado final), una tabla de resúmenes de ratings por nota (número total de valoraciones, número de valoraciones útiles, número de evaluadores únicos y ventanas temporales de evaluación), una tabla agregada de señales por tuit (número de señales, usuarios distintos y momento de la primera señal) y una tabla que cruza señales y notas para medir los retrasos entre la primera señal, la creación de la nota y su eventual aprobación como nota útil.

Sobre estas estructuras se definieron las métricas de análisis. La más relevante es el índice de utilidad de cada nota, calculado como la proporción de valoraciones clasificadas como HELPFUL o SOMEWHAT_HELPFUL sobre el total de valoraciones recibidas, a partir de la variable *helpfulnessLevel*. Además, se calcularon métricas temporales (tiempo entre señal y creación de nota, tiempo entre creación de nota y primera valoración, intervalo entre primera y última valoración, tiempo hasta la aprobación como nota útil) y métricas de participación (número de notas por colaborador, duración de la colaboración, ciclos de *earn-out* y cambios de estado en *userEnrollment*). Finalmente, se aplicó un proceso automático de detección de idioma al texto de cada nota, lo que permitió segmentar los análisis y centrarse específicamente en las notas redactadas en español, comparándolas cuando era necesario con el conjunto de otros idiomas.

El análisis estadístico y visual se realizó íntegramente en R, empleando principalmente funciones del ecosistema *tidyverse* para la limpieza, agregación y visualización de los datos. El uso del paquete *arrow* permitió leer directamente las tablas Parquet almacenadas en S3 y trabajar con particiones remotas sin sobrecargar la memoria local. A partir de estos datos ya modelados se desarrolló un análisis descriptivo que incluye series temporales, distribuciones, comparaciones por tipo de error y tema, indicadores de participación de los colaboradores y patrones de evaluación y utilidad de las notas. El conjunto de datos procesado se guardó en un único archivo listo para su reutilización, y el código empleado se documentará en un repositorio público para facilitar la replicación del estudio¹.

Para identificar específicamente las notas redactadas en español se aplicó un proceso automático de detección de idioma sobre el texto completo de cada anotación. Este procedimiento empleó un modelo de clasificación lingüística basado en *fastText*, ampliamente utilizado en trabajos previos de análisis multilingüe en redes sociales. La detección se realizó nota por nota, asignando el idioma con mayor probabilidad y filtrando aquellas cuyo código correspondía a “es”. A partir de esta operación se observó que, del total de 2.208.038 notas disponibles en el conjunto de datos oficial, 188.868 fueron clasificadas como redactadas en español, lo que representa el 8,55% del total. Esta proporción sitúa al español como el segundo de los idiomas con mayor volumen en el sistema, aunque todavía muy por detrás del inglés (62%) en términos absolutos, y permite acotar el análisis empírico a un subconjunto suficientemente amplio y representativo para estudiar el funcionamiento de las Notas de la Comunidad en el contexto hispanohablante.

4. Resultados

A continuación, para guiar el análisis de los datos obtenidos, los resultados se presentan de acuerdo con las preguntas de investigación formuladas.

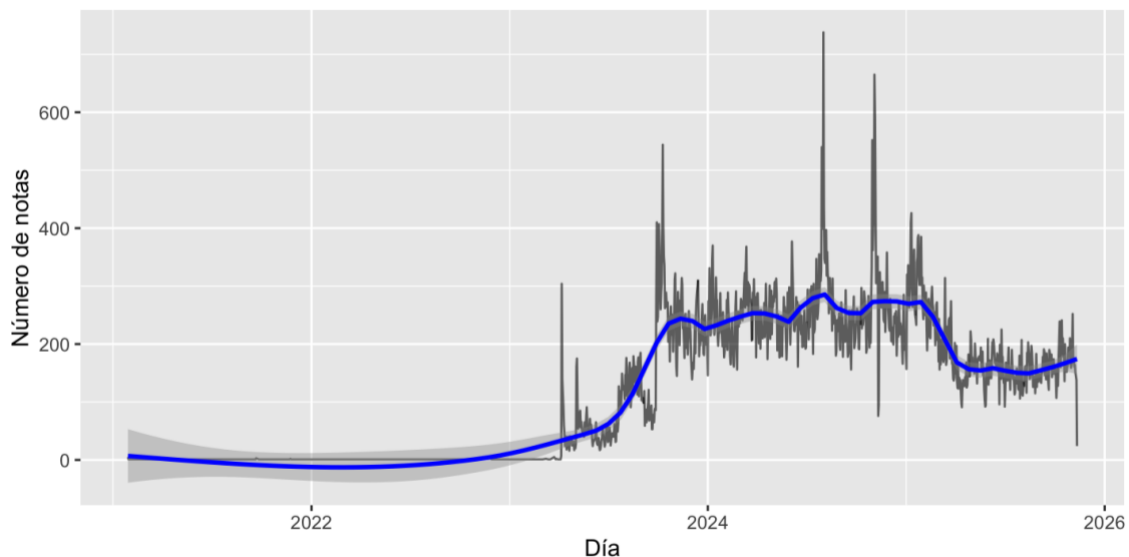
¹ Scripts usados para el procesamiento y análisis de los datos disponibles en: <https://github.com/javicanton/community-notes-analysis>

PII. Frecuencia y dinámica de las notas en español en el sistema de Notas de la Comunidad

La evolución temporal de las notas en español muestra un patrón de crecimiento muy marcado desde la puesta en marcha del sistema hasta el periodo más reciente disponible. En los primeros meses (2021 y buena parte de 2022) el volumen diario de anotaciones es prácticamente residual, con apenas unas pocas notas al día. A partir de 2023 se observa un punto de inflexión claro: el número de notas se incrementa de forma sostenida, primero hasta decenas y luego hasta superar con regularidad los cien registros diarios. La serie diaria presenta picos muy pronunciados (por encima de las 500–600 notas en algunos momentos), que reflejan episodios de alta intensidad informativa y de fuerte activación de la comunidad, previsiblemente asociados a coyunturas políticas, mediáticas o sociales de especial visibilidad.

La línea azul de tendencia suavizada, calculada mediante un ajuste no paramétrico sobre la serie diaria, permite ver mejor esta dinámica general, como muestra la figura 1. La curva se mantiene prácticamente a cero durante la fase inicial del programa, asciende de forma acusada a lo largo de 2023 hasta situarse en torno a las 200–250 notas diarias, y se estabiliza después en un nivel elevado con oscilaciones moderadas. En la parte final del periodo analizado se aprecia un ligero descenso de la tendencia, que sitúa el promedio diario en torno a 150–200 notas, seguido de un repunte, lo que sugiere una fase de reajuste más que un abandono del sistema. De manera que, la combinación de la serie diaria y la tendencia suavizada, indica que las Notas de la Comunidad en español han pasado de ser un mecanismo marginal a constituir un dispositivo de verificación colaborativa, utilizado de forma continuada y con cierta capacidad de respuesta rápida ante momentos de especial conflictividad informativa.

Figura 1. Evolución diaria de Notas de la Comunidad en español, con tendencia suavizada



Fuente: elaboración propia / autores.

Los picos más pronunciados de la serie diaria muestran que la actividad de las Notas de la Comunidad en español se intensifica ante episodios de elevada fricción informativa. Para identificar las causas de estos incrementos, seleccionamos en cada fecha pico los tuits que concentraron más de una nota, lo que permite focalizar el análisis en los contenidos que movilizaron de manera excepcional a la comunidad.

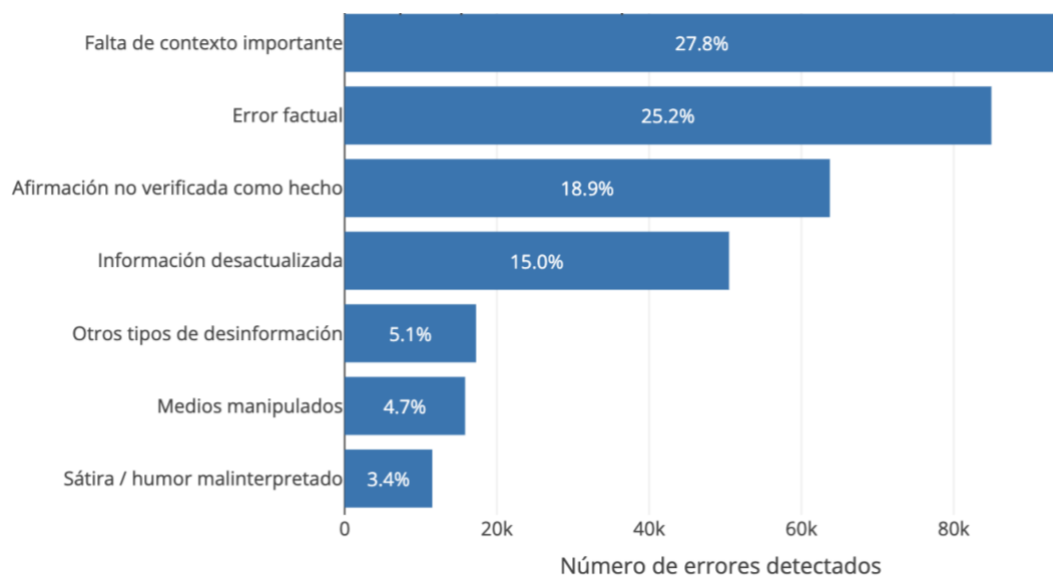
El examen de estos tuits revela una clara concentración temática. Los picos de julio y agosto de 2024 se vinculan con dos controversias de gran visibilidad: las elecciones presidenciales de Venezuela y la difusión internacional de desinformación sobre la boxeadora Imane Khelif durante los Juegos Olímpicos. A finales de octubre y comienzos de noviembre de 2024, el volumen inusualmente alto de anotaciones corresponde a la DANA que afectó a la Comunidad Valenciana, acompañada de una intensa circulación de vídeos descontextualizados, cifras contradictorias y acusaciones políticas. Finalmente, el pico de enero de 2025 coincide con la detención de María Corina Machado en el contexto de la toma de posesión de Nicolás Maduro, un episodio ampliamente mediático que generó olas simultáneas de información manipulada y reacciones polarizadas.

Estos resultados sugieren que los picos de actividad no responden a fluctuaciones aleatorias, sino que reflejan la sensibilidad del sistema a momentos de sobresaturación informativa en los que los colaboradores buscan corregir errores factuales, añadir contexto y contrarrestar narrativas altamente polarizadas.

PI2. Tipos de errores y patrones temáticos en las anotaciones de las Notas en español

El análisis de las notas clasificadas como desinformativas, muestra una distribución claramente concentrada en unas pocas categorías principales. La falta de contexto importante constituye el tipo de error más frecuente, representando el 27,8% del total, seguida de los errores factuales (25,2%) y de las afirmaciones no verificadas como hecho (18,9%). Es importante señalar que una misma nota puede presentar simultáneamente varios tipos de error, por lo que estas proporciones no describen categorías excluyentes, sino la intensidad con la que cada forma de desinformación aparece en el conjunto de anotaciones.

Figura 2. Tipos de error informativo mencionados en las Notas de la Comunidad (español).



Fuente: elaboración propia / autores.

Las categorías restantes presentan un peso significativamente menor. La información desactualizada supone el 15,0%, mientras que otros tipos de desinformación, los medios manipulados y la sátira o humor malinterpretado se mantienen por debajo del 6% cada una. Este patrón indica que los colaboradores hispanohablantes se enfrentan principalmente a contenidos ambiguos, incompletos o imprecisos, más que a materiales profundamente manipulados o de naturaleza satírica. La baja frecuencia relativa de las anotaciones sobre

manipulación audiovisual sugiere que este tipo de verificación (más especializada y exigente) aparece de manera más limitada dentro del sistema.

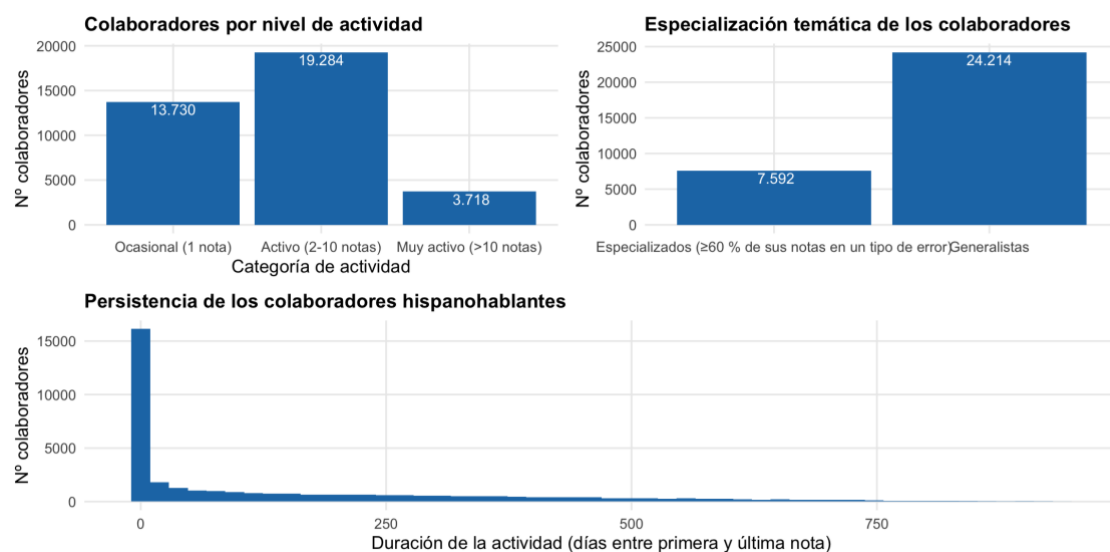
Cuando se examinan los temas que predominan en las notas en español, se observa un patrón complementario a la clasificación de errores. Las anotaciones se concentran sobre todo en ámbitos de alta fricción política y social, destacando las categorías de política, economía, género, migración, seguridad y salud. La política es con diferencia el tema más presente, seguida de la economía y cuestiones vinculadas a género y seguridad. A medida que aumentan las combinaciones temáticas (por ejemplo, política y economía; política, migración y seguridad; o salud, economía y clima) la frecuencia disminuye de forma pronunciada, reflejando una larga cola de temas compuestos menos frecuentes. Este perfil confirma que las notas en español operan principalmente en torno a debates públicos intensos, caracterizados por dinámicas de interpretación conflictiva y alta velocidad de circulación de contenidos.

La tipología observada indica que las Notas de la Comunidad en español funcionan esencialmente como un mecanismo de contextualización y corrección puntual, orientado a frenar la difusión de errores comunes derivados de la circulación acelerada de información incompleta o de afirmaciones presentadas como hechos sin respaldo. La estructura temática de las anotaciones refuerza esta interpretación, mostrando que la comunidad interviene sobre todo en ámbitos donde la disputa interpretativa y la polarización generan un entorno propicio para la desinformación.

PI3. Patrones de participación y enfoque temático de los colaboradores en español

El análisis de los colaboradores hispanohablantes revela un ecosistema amplio, heterogéneo y caracterizado por una fuerte asimetría en los niveles de implicación. Tal como muestra la Figura 3, la mayoría de los participantes contribuye de manera esporádica: un total de 13.730 autores han publicado únicamente una nota, mientras que 19.284 se sitúan en un nivel de actividad intermedio (entre dos y diez notas). Solo un subconjunto reducido (3.718 colaboradores) puede considerarse muy activo, superando las diez notas publicadas. Esta estructura refleja un patrón habitual en sistemas abiertos de colaboración: una base amplia de contribución ocasional y un núcleo operativo pequeño, pero altamente productivo, responsable de buena parte del volumen total de anotaciones.

Figura 3. Patrones de participación de los colaboradores hispanohablantes



Fuente: elaboración propia / autores.

En cuanto a la trayectoria temporal, la persistencia en la actividad presenta también una distribución muy desigual. La mayoría de los colaboradores participa durante períodos extremadamente breves, a menudo limitados al mismo día de su publicación inicial, mientras que un grupo minoritario extiende su actividad durante semanas o meses. Aunque los participantes con trayectorias prolongadas son numéricamente pocos, su presencia sostenida sugiere la existencia de un segmento comprometido que asume funciones de vigilancia continuada, especialmente relevante en momentos de elevada presión informativa.

En términos de orientación temática, los colaboradores muestran un marcado predominio del perfil generalista. Según el umbral establecido, 24.214 autores distribuyen sus contribuciones entre múltiples tipos de error informativo, mientras que únicamente 7.592 pueden considerarse especialistas, es decir, concentran al menos el 60% de sus notas en un mismo tipo de error. Esta configuración sugiere que la colaboración no está impulsada por nichos temáticos cerrados, sino por una intervención transversal ante contenidos variados. Sin embargo, entre quienes sí presentan especialización, los patrones observados en el conjunto de datos indican que los errores dominantes tienden a alinearse con las categorías más frecuentes en el sistema: falta de contexto importante, error factual, afirmaciones no verificadas e información desactualizada, lo cual sugiere que ciertos participantes desarrollan experiencia en corregir tipos de desinformación particularmente recurrentes.

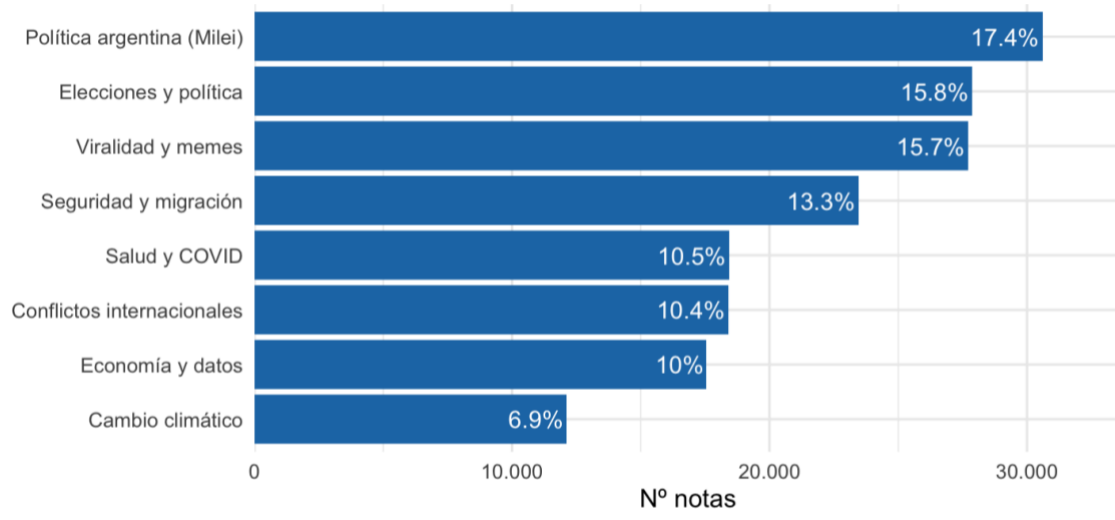
Los datos de participación apuntan a un sistema en el que la contribución hispanohablante combina amplitud y rotación, con un núcleo estable de usuarios activos y moderadamente especializados. Esta estructura híbrida permite una respuesta ágil ante picos de desinformación, al tiempo que sostiene mecanismos mínimos de continuidad y experiencia acumulada.

PI4. Similitud entre las temáticas abordadas por las Notas en español y las áreas informativas cubiertas por los verificadores profesionales

Con el fin de identificar de forma sistemática las principales áreas temáticas abordadas por las notas en español, se aplicó un procedimiento de “detección automática de tópicos” basado en *Latent Dirichlet Allocation* (LDA). Este enfoque permite agrupar miles de anotaciones en función de sus patrones léxicos compartidos, generando conjuntos temáticos coherentes a partir del uso real del lenguaje por parte de los colaboradores. A partir del modelo resultante (configurado con ocho tópicos y evaluado mediante la inspección cualitativa de sus términos más representativos), se elaboró una clasificación temática que sirve como base para analizar en qué medida la actividad comunitaria coincide con las áreas de verificación que desarrollan los equipos profesionales. La Figura 4 sintetiza la distribución final de estos tópicos, expresada en porcentaje relativo.

El análisis temático de las notas en español muestra una clara convergencia con las áreas informativas que los verificadores profesionales tienden a cubrir en sus rutinas de trabajo. La Figura 4 evidencia que los tópicos *Política argentina (Milei)* (17,4%), *Elecciones y política* (15,8%) y *Economía y datos* (10%) concentran una proporción sustancial de las anotaciones generadas por los colaboradores. Esta centralidad de la política y la economía coincide con lo observado en estudios como el de Abuín-Penas et al. (2023), que identifican que la política, la legislación y otros asuntos de actualidad pública constituyen el núcleo temático del fact-checking profesional en España.

Figura 4. Temáticas detectadas mediante LDA en las Notas en español



Fuente: elaboración propia / autores.

En esa misma línea, la presencia destacada de *Salud y COVID* (10,5%), *Seguridad y migración* (13,3%) y *Conflictos internacionales* (10,4%) entre las notas comunitarias refuerza el paralelismo entre ambos modelos de verificación. Tal como muestra el estudio citado, la pandemia, los sucesos internacionales y cuestiones de impacto social fueron áreas de intensa actividad para los verificadores profesionales, lo que se corresponde con la fuerte representación de estos temas en las contribuciones de los usuarios hispanohablantes. De manera que, la distribución observada en la Figura 4 sugiere que las notas comunitarias intervienen en aquellos ámbitos donde la desinformación presenta mayor visibilidad pública y donde los usuarios perciben una necesidad más clara de aportar contexto o corregir afirmaciones.

Sin embargo, el análisis revela también una diferencia significativa: el tópico *Viralidad y memes* (15,7%) adquiere un peso notable en las Notas de la Comunidad, a pesar de que rara vez ocupa un lugar equivalente en las agendas de verificación profesional. Mientras los equipos de fact-checking utilizan plataformas como Instagram para difundir piezas estructuradas y centradas en contenidos de relevancia pública, las notas comunitarias muestran una mayor inclinación a intervenir sobre contenidos virales, humorísticos o ambiguos que generan confusiones rápidas, pero que no suelen justificar una verificación formal. Este comportamiento apunta a que la comunidad cubre un segmento distinto del ecosistema desinformativo, atendiendo a micro-desinformaciones que no forman parte del repertorio habitual de Maldita.es, Newtral, Verificat o Chequeado, pero que sí afectan a la experiencia cotidiana de los usuarios. Este resultado es coherente con investigaciones recientes que muestran la complementariedad existente entre los verificadores profesionales y las Notas de la Comunidad (Fundación Maldita.es, 2025a; Chequeado, 2025).

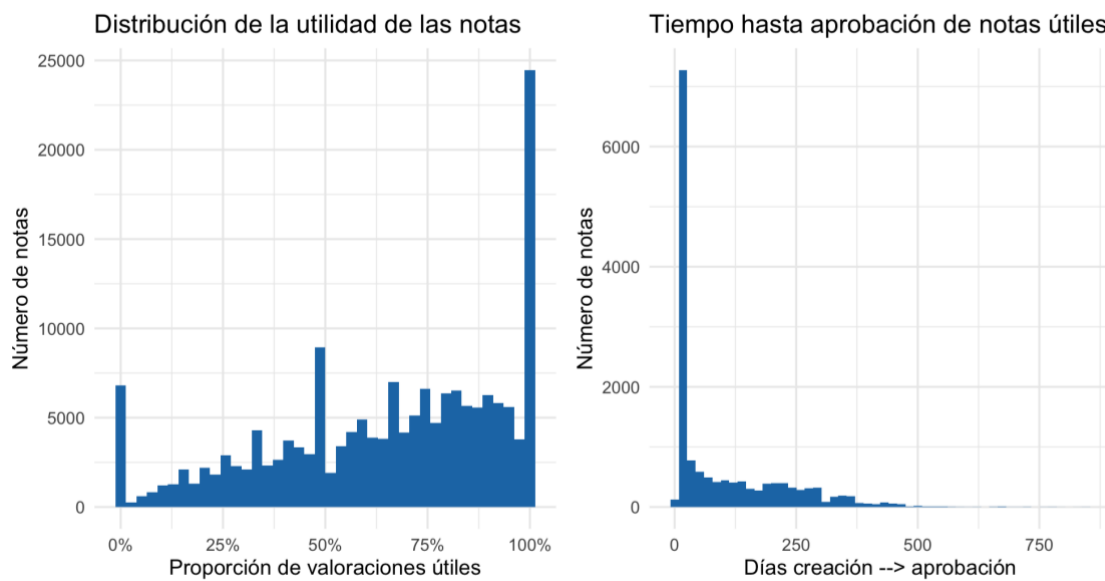
Cabe señalar que, los datos muestran una alta coincidencia estructural entre los temas abordados por las notas en español y las áreas informativas cubiertas por los verificadores profesionales, especialmente en política, economía, salud y asuntos internacionales. Al mismo tiempo, se observa una diversificación funcional propia del sistema colaborativo, que amplía el rango de intervención hacia contenidos virales de baja formalidad que rara vez reciben atención institucional. Este solapamiento parcial sugiere que las Notas de la Comunidad no solo replican estas prioridades profesionales, sino que podrían contribuir a complementar el ecosistema de verificación al aportar cobertura en espacios donde la supervisión editorial es limitada. Esta

complementariedad también ha sido señalada por organizaciones especializadas en verificación digital (Fundación Maldita.es, 2025b).

PI5. Funcionamiento del sistema de consenso en el entorno hispanohablante

El sistema de consenso de las Notas de la Comunidad en español se articula a partir de tres elementos centrales: las valoraciones de utilidad emitidas por los colaboradores, la diversidad estructural de quienes participan en el proceso y los tiempos necesarios para que una nota alcance el estatus de aprobada. El análisis de los datos permite observar un funcionamiento en el que conviven patrones de acuerdo muy robustos con desigualdades significativas en la participación y fricciones temporales en la validación.

Figura 5. Mecanismos de consenso en las Notas en español: distribución de la utilidad de las notas (izquierda) y tiempo en días hasta la aprobación de la nota como útil (derecha)



Fuente: elaboración propia / autores.

En términos de utilidad, las notas en español suelen recibir valoraciones ampliamente favorables. La mediana de la proporción de votos útiles se sitúa en torno al 0,70, lo que indica que, de manera general, los colaboradores reconocen el valor de las notas que aportan contexto. Además, un cuarto de todas las notas supera el 90% de utilidad, lo que señala episodios frecuentes de consenso claro. Las diferencias por tipo de error refuerzan esta tendencia: las notas que corrigen medios manipulados o errores factuales tienden a recibir las valoraciones más elevadas, mientras que aquellas vinculadas a humor o sátira presentan acuerdos algo menores, aunque igualmente altos. Estas variaciones sugieren que el consenso se forma con mayor solidez en torno a desinformaciones claras y verificables.

La participación en forma de votos es, sin embargo, muy desigual. Aunque muchas notas reciben alrededor de una veintena de valoraciones, existe un subconjunto reducido que concentra cientos o miles de votos, con casos extremos que superan los 3.800. Esta concentración indica que, la visibilidad del contenido o la intensidad del debate condicionan la atención que recibe cada nota. No obstante, incluso con esta desigualdad, la utilidad global mantiene un patrón estable, lo que sugiere que el consenso emerge con relativa facilidad cuando la nota aporta información relevante.

Los tiempos de aprobación introducen una dimensión distinta. Si bien, una parte importante de las notas útiles se aprueba en plazos relativamente breves, la mediana se sitúa cerca de un mes. Además, existe una cola larga de casos que requieren varios meses para alcanzar consenso, con algunos que superan el año de espera. Estos retrasos suelen estar asociados a desacuerdos persistentes o a la falta de suficientes valoradores con perspectivas diversas, lo que limita la capacidad del sistema para intervenir de forma oportuna en narrativas de rápida difusión.

La estructura de participación también matiza el funcionamiento del consenso. Aunque la mayoría de cuentas registradas en el sistema corresponden a usuarios nuevos, un 37% de los colaboradores se encuentra en el estado “earnedIn”, que el sistema interpreta como señal de participación estable y fiable. Este grupo sostiene la mayor parte de las aprobaciones y contribuye a la coherencia de los consensos. Sin embargo, esta dependencia también puede reducir la diversidad efectiva del proceso, especialmente cuando las notas implican interpretaciones controvertidas o temas polarizados.

En definitiva, el sistema de consenso en español parece caracterizarse por acuerdos generalmente sólidos, una comunidad estable de colaboradores fiables y una clara capacidad para distinguir entre tipos de correcciones. Al mismo tiempo, la desigual participación evaluadora y los tiempos prolongados en un subconjunto significativo de notas señalan los límites estructurales del sistema, especialmente en contextos que requieren intervenciones rápidas o mayor pluralidad de perspectivas.

5. Discusión

Los resultados obtenidos permiten comprender en profundidad el funcionamiento del sistema de Notas de la Comunidad en español y evaluar en qué medida este reproduce, amplifica o transforma las dinámicas descritas en estudios previos sobre verificación colaborativa y moderación algorítmica.

En relación con la PI.1, los resultados muestran un incremento notable en el volumen de verificaciones mediante Notas de la Comunidad, especialmente en tuits publicados por usuarios verificados con un elevado número de seguidores. Sin embargo, este aumento en la actividad de verificación no se traduce directamente en una reducción significativa de la interacción con contenidos potencialmente engañosos en X. Este hallazgo es coherente con lo señalado por Chuai et al. (2024), quienes advierten de la limitada capacidad de los mecanismos de verificación colaborativa para modificar de forma sustancial los patrones de *engagement* con la desinformación. Este hallazgo nos plantea la cuestión de si es suficiente aumentar la cantidad de verificaciones si estas no logran frenar la viralidad inicial de la desinformación, siendo uno de los principales riesgos que hay que tratar de paliar con estos contenidos. Ya que los resultados obtenidos muestran que, si bien las notas de la comunidad se trata de un sistema que ha crecido rápidamente, su función principal es corregir errores informativos de relativa baja complejidad epistémica y depende de una comunidad pequeña de colaboradores activos, unos 3.700 usuarios muy activos según nuestros datos, lo cual hace que la desinformación vaya a un ritmo más rápido que la propia verificación, pero que parece suficiente para el fundador de Meta, Mark Zuckerberg, quien afirmaba en enero de 2025: “Vamos a volver a nuestras raíces y centrarnos en reducir errores, simplificar nuestras políticas y restaurar la libertad de expresión en nuestras plataformas”, afirmó en un vídeo (Meta, 2025).

Así, en la PI.2 se identifican tensiones importantes derivadas del diseño algorítmico basado en consenso transversal, especialmente en contextos de fuerte polarización temática o política. El algoritmo de Notas de la Comunidad, conocido como “bridging algorithm”, destaca aquellas notas que son valoradas como útiles por usuarios con perspectivas diversas. En lugar de basarse en la

cantidad total de votos positivos, el sistema identifica cuándo una nota recibe apoyo simultáneo de personas con historiales de evaluación distintos, incluidas posiciones opuestas del espectro político, y solo en esos casos la publica y la muestra al resto de usuarios (Wojcik et al., 2022), lo que coincide con estudios que relacionan la valoración con el alineamiento político percibido entre evaluadores y autores, esta dinámica de polarización podría explicar por qué un elevado porcentaje de notas nunca alcanza visibilidad pública.

Al mismo tiempo, en la PI.3, los resultados evidencian que la participación en la evaluación de notas no es homogénea, puesto que el volumen desigual de ratings sugiere que factores externos, como la visibilidad del tuit original (número de impresiones, viralidad) y su relevancia temática (interés social o polarización), influyen directamente en la disposición de los usuarios a calificar. Esto implica que el modelo colaborativo no opera en condiciones neutrales, y que los usuarios tienen un fuerte interés en desacreditar la desinformación política (Pröllochs, 2022).

Además, la fuerte asimetría en los tiempos de aprobación, con la mayoría de notas validadas en pocas horas y una minoría que requiere días o semanas, plantea desafíos sobre la funcionalidad del sistema. Según apunta un estudio realizado Instituto de Democracia Digital de las Américas la gran mayoría de las Notas de la Comunidad en X presentadas por los usuarios nunca se publican, suponiendo únicamente alrededor de un 10% de todas las recibidas (DDIA, 2025). La rapidez es crucial para limitar la viralización de contenidos engañosos. Numerosos estudios muestran que la desinformación tiende a circular más rápido y más ampliamente que los contenidos verificados, lo que sitúa a cualquier intervención correctiva en una posición de desventaja si llega demasiado tarde (Vosoughi et al., 2018) y su capacidad para contrarrestar la desinformación disminuye significativamente, ya que la mayor parte de la audiencia ya ha estado expuesta al contenido engañoso (Li & Chang, 2023). Un sistema de consenso multidimensional, diseñado para garantizar diversidad y evitar sesgos partidistas, puede generar un cuello de botella temporal que impida que notas correctas lleguen al público en el momento más relevante de la conversación, como reflejan los tiempos de aprobación observados en nuestros datos, con una mediana cercana a los 27 días.

La evolución temporal observada en nuestro conjunto de datos muestra un crecimiento sostenido del uso de las Notas de la Comunidad en español durante la mayor parte del periodo analizado, lo que sugiere su progresiva consolidación como mecanismo de moderación participativa en el ecosistema digital, si bien se aprecia un descenso en su uso durante el último año. Asimismo, los incrementos de actividad tienden a coincidir con periodos de alta conflictividad informativa o politización, un patrón ampliamente documentado: en contextos de crisis o incertidumbre, la desinformación se intensifica al generar mayor activación emocional, lo que impulsa la movilización colectiva para su detección y corrección (Morejón-Llamas et al., 2025). Este comportamiento es similar al observado en los comentarios de medios tradicionales, donde la participación ciudadana se orienta a influir en la esfera pública, expresar percepciones sobre situaciones de crisis, fiscalizar a actores institucionales y ampliar el debate más allá del contenido de las noticias (Molina Rodríguez, 2022). Esta coincidencia contextual sugiere que los usuarios pueden actuar como una forma de “detección distribuida” de contenidos problemáticos, reaccionando de manera más intensa ante situaciones percibidas como informativamente críticas. No obstante, este impulso inicial no siempre se traduce en una validación inmediata de las notas, ya que los mecanismos de consenso entre evaluadores, la verificación de evidencias y la propia dinámica deliberativa del sistema pueden introducir retrasos, lo que potencialmente reduce la capacidad de respuesta en momentos de alta demanda informativa.

En relación a la PI.4, los resultados muestran que las notas generadas por la comunidad se enfocan principalmente en aportar contexto y corregir errores factuales o afirmaciones no verificadas, en lugar de identificar manipulación técnica o contenido satírico, tareas que requieren

competencias especializadas. Esto sugiere que el sistema funciona como una herramienta de contextualización y mejora de la precisión informativa. Además, se observa que la mayoría de las notas se concentran en temáticas de alto impacto público, como política, economía, seguridad e inmigración, lo que refleja una priorización de los ámbitos en los que la desinformación tiene mayores implicaciones sociales. Complementariamente, estudios empíricos recientes indican que un factor determinante de la utilidad percibida de las verificaciones comunitarias es la inclusión de enlaces a fuentes externas que respalden las afirmaciones realizadas por los usuarios. De hecho, se ha encontrado que la probabilidad de que una verificación comunitaria sea considerada útil es, en promedio, 2,33 veces mayor cuando contiene referencias a fuentes externas (Drolsbach et al., 2025). Asimismo, Pröllochs (2022) señala que las verificaciones comunitarias tienden a ser más útiles cuando incorporan emociones positivas y evitan el lenguaje incendiario o provocador.

Estos hallazgos sugieren que la capacidad de proporcionar evidencia verificable constituye un componente central para maximizar la efectividad de la verificación basada en la comunidad y resalta la importancia de incentivar buenas prácticas de citación en este tipo de plataformas.

Lo relativo a la PI.5, el patrón de participación (con un núcleo relativamente pequeño de usuarios muy activos y una mayoría amplia de colaboradores ocasionales) sigue una distribución de tipo ley de potencia, una estructura recurrente en comunidades colaborativas de producción voluntaria, como Wikipedia o Reddit (Ruprecht et al., 2021). La presencia de colaboradores con larga trayectoria sugiere cierto grado de institucionalización informal, aunque el peso excesivo de unos pocos contribuyentes plantea interrogantes sobre la diversidad real de perspectivas, un elemento esencial para el algoritmo de consenso. La especialización limitada, con predominio de usuarios generalistas, indica que el sistema depende de contribuyentes capaces de aportar contexto general y verificar hechos básicos, reforzando la interpretación del sistema como una herramienta de inteligencia colectiva más que como un mecanismo experto. Sin embargo, la baja presencia de expertos podría limitar la eficacia en categorías más complejas o sensibles.

6. Conclusiones

Los hallazgos apuntan a un sistema de verificación colaborativa que funciona como mecanismo complementario y parcialmente dependiente del fact-checking profesional, con eficacia desigual según el tipo de error, el tema y el grado de polarización. La estructura comunitaria reproduce dinámicas conocidas de participación en línea, con alta concentración y fuerte desigualdad, mientras que el algoritmo de consenso introduce oportunidades y limitaciones, especialmente en términos de rapidez y diversidad. El caso del español aporta evidencia original sobre la evolución del sistema fuera del contexto anglosajón, permitiendo evaluar de forma más matizada las posibilidades y límites de la verificación distribuida.

Este estudio presenta algunas limitaciones. Se centra únicamente en el entorno hispanohablante de X, por lo que los hallazgos podrían no ser generalizables a otros idiomas o contextos culturales. Además, se basa en la actividad observable en la plataforma, sin considerar factores externos que podrían influir en la creación o evaluación de notas, ni posibles sesgos derivados de la categorización de errores y de la valoración comunitaria. La temporalidad del análisis limita la observación de cambios a largo plazo en la comunidad y en la dinámica del sistema.

A partir de estos resultados, se identifican varias líneas de investigación futura. Sería relevante explorar cómo la colaboración con verificadores profesionales puede optimizar la eficacia en errores complejos y en contextos polarizados. También conviene estudiar el impacto de la diversidad de la comunidad en la calidad del consenso y los tiempos de aprobación, así como la relación entre visibilidad del contenido y efectividad de las correcciones. Por último, comparar

la dinámica del sistema en distintos idiomas y contextos culturales permitiría identificar patrones universales y diferencias contextuales, contribuyendo al diseño de estrategias más efectivas contra la desinformación.

Reconocimientos

Esta publicación es parte del proyecto MonitorIA: herramienta avanzada para la monitorización y detección temprana de desinformación en Telegram mediante IA (T25-003-I) financiada por la convocatoria del Programa Propio de Investigaciones Estratégicas para el Desarrollo de Capacidades de Investigación y Transferencia de la Universidad Internacional de la Rioja (Programa Incentiva 25).

References

- Abuín-Penas, J., Corbacho-Valencia, J. M., & Pérez-Seoane, J. (2023). *Análisis de los contenidos verificados por los fact-checkers españoles en Instagram*. *Revista de Comunicación*, 22(1), 17–34. <https://doi.org/10.26441/RC22.1-2023-3089>
- Allen, J., Martel, C., & Rand, D. G. (2022). Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. *CHI Conference on Human Factors in Computing Systems*, 1–19. <https://doi.org/10.1145/3491102.3502040>
- Allen, J., Arechar, A. A., Pennycook, G., & Rand, D. G. (2021). Scaling up fact-checking using the wisdom of crowds. *Science Advances*, 7(36), eabf4393. <https://doi.org/10.1126/sciadv.abf4393>
- Amazeen, M. A. (2020). Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, 21(1), 95–111. <https://doi.org/10.1177/1464884917730217>
- Cantón-Correa, J. (2024). *Disinformation as a global threat: A sociology of risk analysis*. In R. Gelado Marcos, J. Sixto-García, & J. E. González Vallés (Coords.), *Disinformation and risk society: What is at stake with the proliferation of information disorders*. Fragua.
- Cantón-Correa, J., Montoro-Montarroso, A., Gómez-Romero, J., & Molina-Solana, M. (2025a). Perfiles y necesidades formativas de los fact-checkers de la Península Ibérica: impacto de la IA. *Doxa Comunicación. Revista Interdisciplinar De Estudios De Comunicación Y Ciencias Sociales*, 40, 465–491. <https://doi.org/10.31921/doxacom.n40a2725>
- Cantón-Correa, F.-J., Ballesteros-Aguayo, L., & Montoro-Montarroso, A. (2025b). A fact-checking tool based on Artificial Intelligence to fight disinformation on Telegram. *Communication & Society*, 38(1), 247–262. <https://doi.org/10.15581/003.38.1.019>
- Chequeado. (2025, febrero 14). Notas de la Comunidad en X: los verificadores de datos independientes son una de las fuentes más citadas y útiles para los usuarios. <https://chequeado.com/el-explicador/notas-de-la-comunidad-en-x-los-verificadores-de-datos-independientes-son-una-de-las-fuentes-mas-citadas-y-utiles-para-los-usuarios/>
- Chuai, Y., Tian, H., Pröllochs, N., & Lenzini, G. (2024). *Did the roll-out of Community Notes reduce engagement with misinformation on X/Twitter?* Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), Article 428, 1–52. <https://doi.org/10.1145/3686967>
- Dafonte-Gómez, A., Corbacho-Valencia, J. M., & García-Mirón, S. (2021). El fact-checking en Iberoamérica: Evolución reciente y mapa de situación. En J. Sotelo González & J. González García (Eds.), *Digital Media. El papel de las redes sociales en el ecosistema comunicativo en tiempo de COVID-19* (pp. 877–889). McGraw-Hill.
- Digital Democracy Institute of the Americas (DDIA). (2025). A Deep Dive into X's Community Notes: An Analysis of English and Spanish Contributions Between 2021 and 2025. Digital Democracy Institute of the Americas.

- Drolsbach, C. P., Solovev, K., & Pröllochs, N. (2024). Community notes increase trust in fact-checking on social media. *PNAS Nexus*, 3(7) <https://doi.org/10.1093/pnasnexus/pgae217>
- European Commission. (2023). New call for proposals supports EU fact-checkers in debunking disinformation. <https://digital-strategy.ec.europa.eu/en/news/new-call-proposals-supports-eu-fact-checkers-debunking-disinformation>
- Fundación Maldita.es. (2025a). Faster and more useful: The impact of fact-checkers in X's Community Notes. <https://maldita.es/investigaciones/20250213/community-notes-factcheckers-impact-report/>
- Fundación Maldita.es. (2025b, enero 17). Por qué las Notas de la Comunidad y los verificadores pueden (y deben) trabajar juntos, y cómo hacerlo para que funcione. <https://maldita.es/nosotros/20250117/notas-comunidad-verificadores-juntos-funcione/>
- Google. (2025). Letter from Kent Walker to the European Commission regarding the EU Code of Practice on Disinformation. Google.
- Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.
- Kyriakidou, M., Cushion, S., Hughes, C., & Morani, M. (2022). Questioning fact-checking in the fight against disinformation: An audience perspective. *Journalism Practice*, 17(10), 2123–2139. <https://doi.org/10.1080/17512786.2022.2097118>
- Li, J., & Chang, X. (2023). Combating misinformation by sharing the truth: A study on the spread of fact-checks on social media. *Information Systems Frontiers*, 25, 1479–1493. <https://doi.org/10.1007/s10796-022-10296-z>
- Martel, C., Allen, J., Pennycook, G., & Rand, D. G. (2024). Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science*, 19(2), 477–488. <https://doi.org/10.1177/17456916231190388>
- Masotina, M., Musi, E., & Yates, S. (2024). Relevance theory for mapping cognitive biases in fact-checking: An argumentative approach. *Frontiers in Psychology*, 15, 1468879. <https://doi.org/10.3389/fpsyg.2024.1468879>
- Meta. (2025). More speech and fewer mistakes.
- Molina Rodríguez, L. (2022). Formas de participación de los usuarios de la prensa digital. Cubadebate como caso de estudio. *Cuadernos.Info*, (52), 160–180. <https://doi.org/10.7764/cdi.52.31283>
- Montoro-Montarroso, A., Cantón-Correa, J., Rosso, P., Chulvi, B., Panizo-Lledot, Á., Huertas-Tato, J., Calvo-Figueras, B., Rementeria, M. J., & Gómez-Romero, J. (2023). Fighting disinformation with artificial intelligence: Fundamentals, advances and challenges. *Profesional de la información*, 32(3), e320322. <https://doi.org/10.3145/epi.2023.may.22>
- Morejón-Llamas, N., & Tarín Sanz, A. (2025). Fact-checking y cooperación transnacional: Análisis de la respuesta europea a la desinformación bélica en el contexto de las Elecciones Europeas de 2024. *Revista de Comunicación*, 24(2), 297–320. <https://doi.org/10.26441/rc24.2-2025-3858>
- Moreno-Gil, V., & Salgado-de Dios, F. (2023). El cumplimiento del código de principios de la International Fact-Checking Network en las plataformas de verificación españolas: Un análisis cualitativo. *Revista de Comunicación*, 22(1), 293–307. <https://doi.org/10.26441/RC22.1-2023-2971>
- Nyariki, E. (2025). *The State of the Fact-Checkers Report*. International Fact-Checking Network, The Poynter Institute.
- Poch-Butler, S. L., Gelado-Marcos, R., Ventura-Salom, B., & De la Calle Velasco, G. (2024). Fact-checking en procesos electorales y campaña permanente. Un análisis comparado entre España y Portugal. *Revista Latina De Comunicación Social*, (83), 1–27. <https://doi.org/10.4185/rlds-2025-2318>

- Pröllochs, N. (2022). Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 16, pp. 794–805). <https://ojs.aaai.org/index.php/ICWSM/article/view/19335>
- Reuters Institute. (2025). *Digital news report 2025*. University of Oxford
- Rupprechter, T., Horta Ribeiro, M., Santos, T., Lemmerich, F., Strohmaier, M., West, R., & Helic, D. (2021). Volunteer contributions to Wikipedia increased during COVID-19 mobility restrictions. *Scientific reports*, 11(1), 21505. <https://arxiv.org/abs/2102.10090>
- Singer, J. (2021). Border patrol: The rise and role of fact-checkers and their challenge to journalists' normative boundaries. *Journalism: Theory, Practice and Criticism*, 22(8), 1929–1946. <https://doi.org/10.1177/1464884920933137>
- Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). Community Notes reduce engagement with and diffusion of false information online. *Proceedings of the National Academy of Sciences*, 122(38), e2503413122. <https://doi.org/10.1073/pnas.2503413122>
- TikTok. (2025). Testing a new feature to enhance content on TikTok.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research*, 47(2), 155–177. <https://doi.org/10.1177/0093650219854600>
- Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe report*. <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>
- Westlund, O., Belair-Gagnon, V., Graves, L., Larsen, R., & Steensen, S. (2024). What Is the Problem with Misinformation? Fact-checking as a Sociotechnical and Problem-Solving Practice. *Journalism Studies*, 25(8), 898–918. <https://doi.org/10.1080/1461670X.2024.2357316>
- Wojcik, S., Hilgard, S., Judd, N., Mocanu, D., Ragain, S., Hunzaker, M. B. F., Coleman, K., & Baxter, J. (2022). Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation. *arXiv*. <https://doi.org/10.48550/arXiv.2210.15723>