Medir la corroboración, no la veracidad: un índice de confianza con pruebas para la prensa digital

Óscar Trabazos Clavería · Working paper, sin revisión por pares · Versión 1.0, 2 de noviembre de 2026

Resumen

Este documento propone medir cuánta corroboración tiene una información en la prensa digital, en lugar de declararla verdadera o falsa. Describe GeriAI, un sistema que, a partir de un titular o un enlace, busca piezas que cuentan lo mismo, agrupa las copias de una misma fuente, decide si cada pieza apoya, contradice o no dice nada, y combina nueve señales en un índice de 0 a 100 que siempre llega con sus pruebas.

El trabajo recoge una línea que empezó en 2007 midiendo el tono y la reputación en lo que se publica, que en 2017 se preguntó «¿cómo detectar una noticia falsa?» y que en 2026 se reformuló como «¿cuánta confianza merece y con qué pruebas?». Presenta el método, la traducción de una propuesta sin medir de 2023 (la «técnica del impostor») a un sistema medido, y una evaluación preliminar: unos diez casos reales y una prueba de clasificación sobre 251 pares de titulares.

La evaluación es preliminar y se dice así: los pesos son provisionales y no están calibrados, los casos son ejemplos y no una muestra, y parte de las etiquetas de la prueba las puso una IA sin revisión humana todavía. El documento separa lo hecho y medido de lo propuesto y de lo que queda abierto.

Palabras clave: desinformación, corroboración, fuentes independientes, inferencia textual, índice de confianza, prensa digital.

Sobre el autor

Óscar Trabazos Clavería lleva desde 2007 construyendo sistemas que leen lo que se publica en prensa e internet y miden algo más que palabras: primero el tono y la reputación de las marcas, después la ironía y la emoción, y desde 2017 la desinformación.

- **Anpro21 y Brand Rain.** En Anpro21 lanzó en 2009 una medida del tono de las marcas en prensa e internet, y en 2010 Brand Rain, una herramienta de medición de reputación. Su base técnica se publicó en la revista *Procesamiento del Lenguaje Natural* de la SEPLN (2014).

- **TrawlingWeb.** Es su fundador y consejero delegado. TrawlingWeb analiza prensa digital y redes sociales de decenas de países, y GeriAI es su capa de análisis con IA. El sistema de este documento trabaja sobre ese corpus.
- **Investigación con universidades.** Es coautor, con Pilar Sánchez-García (Universidad de Valladolid) y João Canavilhas (Universidade da Beira Interior), de un estudio sobre cómo la prensa de España y Portugal cuenta la IA, hoy en evaluación. En septiembre de 2026 dio la conferencia inaugural del I Congreso Internacional IBERPERIA (Covilhã).
- **Escritura.** Publica en su blog desde 2006; la serie sobre desinformación va de 2017 a 2026.

Este working paper es la primera vez que la línea sobre desinformación se presenta con método y mediciones. Lo anterior fueron textos de divulgación, propuestas y prototipos, y así se trata aquí.

1. Introducción

La pregunta de este trabajo es cuánta confianza merece una información publicada, medida por cuánto la corroboran fuentes independientes de la prensa digital en un periodo dado. No pregunta si es verdadera.

Los engaños no son nuevos. Lo nuevo es la velocidad: cualquiera publica, nadie tiene que contrastar antes y lo que encaja con lo que el público quiere oír se multiplica solo. La respuesta clásica del periodismo sigue valiendo: contrastar con fuentes independientes antes de dar algo por bueno. Lo que cambia es la escala a la que hay que hacerlo.

Un veredicto binario falla en los dos sentidos. Un hecho cierto pero reciente o local puede tener poca cobertura; un bulo muy copiado puede tener mucha. Por eso el sistema no dice «verdadero» ni «falso»: da un índice de 0 a 100, la etiqueta de su tramo (de «sin corroboración suficiente» a «corroboración alta») y todas las piezas que lo sostienen o lo contradicen, para que juzgue quien lee.

Este documento aporta tres cosas:

1. Un método explícito para medir corroboración: cómo se buscan las piezas, cómo se cuentan las fuentes independientes, cómo se decide si apoyan o contradicen y cómo se combinan las señales.
2. El recorrido de una propuesta metodológica de 2023, que no tenía cifras, a un sistema en producción con cada cambio medido en casos reales.
3. Una evaluación preliminar con sus límites a la vista.

No aporta una evaluación sistemática: no hay un conjunto de prueba amplio con etiquetas revisadas por personas ni una comparación con verificadores profesionales. La sección 7 detalla qué se puede y qué no se puede concluir.

2. Antecedentes y marco

El trabajo se apoya en cuatro ideas de la literatura y en una regla del oficio. No es una revisión sistemática: cita lo que ha orientado el diseño.

- **Tipos de desorden informativo.** Wardle y Derakhshan (2017) distinguen información errónea, desinformación y malinformación, y siete formas: sátira, conexión falsa, contenido engañoso, contexto falso, contenido impostor, contenido manipulado y contenido inventado. Este marco describe el *tipo* de engaño; el índice de este trabajo no lo clasifica, mide otra cosa (la corroboración).
- **Motivación del emisor.** Las «cuatro P» que Eliot Higgins atribuye a la desinformación (pasión, política, propaganda y pago), recogidas por Wardle (2017), explican por qué se fabrica. No se miden aquí: quedan como contexto.
- **El texto solo no basta.** Shu et al. (2017) revisan la detección automática en redes y concluyen que el contenido debe completarse con el contexto social: fuentes, red y difusión. Tacchini et al. (2017) clasifican publicaciones por quién interactúa con ellas, no por lo que dicen. Ahmed (2017) muestra el alcance y el límite de los rasgos léxicos (n-gramas). Este trabajo toma la conclusión común: lo que delata una información está sobre todo fuera de su texto.
- **Verificación de afirmaciones con pruebas.** La línea de verificación automática de afirmaciones (por ejemplo, FEVER, Thorne et al., 2018) recupera textos y decide si apoyan o refutan una afirmación, con la inferencia textual (Bowman et al., 2015) como pieza central. GeriAI usa ese esquema, pero sobre prensa del mismo periodo y para medir corroboración, no para emitir un veredicto.
- **La regla más antigua del periodismo.** Contrastar con fuentes independientes antes de publicar. El índice la convierte en números: cuántas fuentes distintas, de cuántos países, cuántas apoyan y cuántas contradicen.

La diferencia con la mayor parte de la literatura de detección es el objeto de la medida. No se entrena un clasificador de «falso/verdadero»; se cuenta y se cualifica la corroboración disponible, y se enseña.

3. Trayectoria de la línea (2007-2026)

La pregunta cambió tres veces: de entender el tono y la reputación (2007-2014) a detectar lo falso (2017-2023) y de ahí a medir la confianza (2025-2026). En desinformación, hasta 2026 todo fueron textos de divulgación, propuestas y prototipos; las primeras mediciones son las de este documento.

Año	Hito	Qué aporta a este trabajo	Tipo de fuente
2007-2009	En Anpro21, primer intento de un algoritmo que entendiera el sentimiento de lo publicado; en 2009, una medida del «tono» de cada marca en prensa e internet	La pregunta de fondo: medir a escala lo que se publica. Y el primer muro, el sarcasmo, que una máquina de entonces no percibía	Blog (2009)
2010-2011	Brand Rain: reputación y percepción de marca; estudio de las elecciones generales del 20N con unos 15.000 mensajes en tres semanas	Medir a la vez prensa, blogs y redes, y aplicarlo a la política	Blog (2011)
2012	Proyecto de detección de amenazas, según el autor por encargo del Ministerio del Interior a través de una empresa socia que desarrollaba software de defensa	Según el autor, el origen de la idea de «detectar al impostor»	Relato del autor, sin documento
2013	Ironía, sarcasmo y tono en la medición de intangibles	La ironía como el error más visible de cualquier medición automática de opinión	Blog (2013)
2014	Artículo en *Procesamiento del Lenguaje Natural* n.º 53 (SEPLN), sobre PLN y tecnología semántica en Brand Rain y Anpro21	La base técnica: escucha activa, entidades y relaciones en prensa. No trata de desinformación	Publicación revisada
2014	Proyecto de video-reputación con Ibermática, Verbio, CSA y el grupo BISITE de la Universidad de Salamanca, con apoyo del Plan Avanza	Si el texto y la imagen no cuadran, hay que mirar dos veces: la idea vuelve en 2024 con el vídeo sintético	Blog (2014)
2017	«Las fake news en la política: un desafío para la democracia» (blog)	El problema es la difusión, no el engaño en sí	Divulgación

Año	Hito	Qué aporta a este trabajo	Tipo de fuente
2017-2020	Jimcrick, aplicación contra el acoso escolar de la sociedad NAD	Detectar amenazas en las conversaciones de menores; según el autor, con la misma idea del impostor	Documentos de NAD (2018-2020)
2018	Serie «Políticos, fake news y big data» (2 partes, inédita hasta ahora; firmada entonces con seudónimo)	Presenta el algoritmo de Sirajudeen et al. (2017): comprobar la fuente y buscar si el título y el tema aparecen en otros sitios. Es la primera vez que la corroboración externa aparece en la línea, como método ajeno	Borrador propio sobre trabajo ajeno
2023	Serie «Detectar fake news» I-III: titulares semánticos, clasificación en dos etapas, Wardle y Higgins	El marco teórico y la técnica del impostor (apartado 3.1)	Divulgación + abstract inédito
2023	Prototipo conversacional sobre el corpus de TrawlingWeb, probado con el despido de Sam Altman	Buscaba el titular en 12 idiomas y puntuaba con estrellas. Sin código de motor ni mediciones	Prototipo + vídeos de demostración
2024	«Sora: cambiando las reglas del juego de la desinformación»	La coherencia entre texto, imagen y emoción como señal frente al vídeo sintético	Divulgación
2025	«Seguridad temática» para quien redacta	Tres ejes: alcance, credibilidad («¿quién más habla del tema?») y autoridad del emisor. El segundo eje es el índice de este trabajo	Divulgación
2026	Serie sobre desinformación como riesgo y verificación con modelos generativos	El giro: «no buscamos la verdad»; patrones de propagación (velocidad frente a volumen, homogeneidad, historial de la fuente)	Divulgación

Año	Hito	Qué aporta a este trabajo	Tipo de fuente
2026	GeriAI: índice de corroboración en producción, con registro de mediciones	El sistema y la evaluación de las secciones 4 a 6	Sistema + registro de mediciones

Dos avisos sobre esta trayectoria. Primero, algunos textos de divulgación de 2023 prometían más de lo que existía (validar «en tres minutos» si algo era falso, redes adversarias, cadena de bloques); nada de eso se construyó y este documento no lo recoge como resultado. Segundo, el abstract de la técnica del impostor (2023) sitúa su origen en un proyecto de detección de amenazas de 2012 (según el autor, un encargo del Ministerio del Interior a través de una empresa socia que desarrollaba software de defensa; el abstract lo atribuía por error a la OTAN) y en su uso posterior en Jimcrick, una aplicación contra el acoso escolar. Que Jimcrick existió y era una aplicación contra el acoso consta en documentos de 2018 a 2020; que usara esta técnica, y el proyecto de 2012, no tienen documento. Por eso figuran en la tabla como relato del autor y no sostienen ninguna conclusión de este trabajo.

3.1 La técnica del impostor (2023)

La técnica del impostor propone juzgar una información por quién la emite y por dónde empezó, no solo por lo que dice. Se formuló en un abstract de noviembre de 2023 que no llegó a publicarse y se divulgó en la tercera parte de la serie «Detectar fake news».

La idea de partida es interrogar a todos los que participan en una conversación (usuarios, periodistas, medios, autoridades) para descubrir quién no es lo que dice ser: el impostor. Según el autor, nació en el proyecto de detección de amenazas de 2012 y se reutilizó contra el acoso escolar. El abstract la aplica a la desinformación con tres componentes:

1. **Zona 0.** Rastrear hacia atrás hasta el instante, el lugar y la fuente que trataron el tema por primera vez. Quien lo contó primero y cómo se extendió desde ahí dicen más que el texto de cualquier copia.
2. **Autoridad del emisor.** Revisar el historial del medio y del autor que firma: si ya trataban ese tema, con qué consistencia y con qué reputación en ese ámbito.
3. **IA discriminativa e IA adversaria.** Una IA clasifica y otra cuestiona lo que la primera concluye, para reducir los falsos positivos.

El marco teórico combina a Wardle, para saber *qué tipo* de engaño hay, y a Higgins, para preguntarse *por qué* se difunde (pasión, política, propaganda o pago). La autoridad del emisor es la respuesta práctica a una de las formas de Wardle, el contenido impostor, que da nombre a la técnica.

Su límite es que era una propuesta sin medir. El abstract afirma que la técnica «ha mostrado ser una herramienta eficaz», pero no da ninguna cifra; su sección de resultados enumera líneas futuras y la de referencias quedó sin escribir. El propio texto reconoce que no se había comparado con otras técnicas. Este trabajo toma las tres ideas y las somete a medición (sección 5).

3.2 Qué tipos de Wardle puede ver la corroboración

La corroboración ve bien el contenido inventado y muy poco del resto. Medir cuánto se confirma un hecho no dice, por sí solo, de qué forma engaña una pieza.

Tipo (Wardle y Derakhshan, 2017)	Qué puede ver el índice	Por qué
Contenido inventado	Sí	Sin fuentes independientes que lo cuenten, el índice no pasa de 25 (control inventado y fuente única, sección 6)
Contenido engañoso	En parte	Las reglas de cifra distinta, atribución y cautela evitan contar como apoyo lo que no lo es; no etiquetan el engaño
Conexión falsa	En parte	En cada coincidencia se evalúan título y cuerpo por separado; que el titular analizado no resuma su propio cuerpo no se comprueba
Contenido impostor	Poco	Un medio suplantado que publica solo queda en «una sola fuente»; detectarlo de verdad exige la autoridad del emisor, que no está hecha
Contexto falso	No	Si el hecho es cierto, se corrobora aunque se presente en un contexto falso
Contenido manipulado	No	El sistema solo trata texto; no analiza imágenes ni vídeo
Sátira o parodia	No	Puede salir sin corroboración, pero el sistema no sabe que es sátira

Por eso el índice no sustituye a la clasificación de Wardle; la complementa. El tipo de engaño podría añadirse como una etiqueta aparte, sin mezclarla con el número (sección 8).

4. Método

El índice es una suma ponderada de nueve señales de corroboración, recortada por cuatro reglas fijas y acompañada de todas las piezas que la sostienen. Cada umbral y cada regla se fijó midiendo casos reales; los pesos, no.

4.1 Preguntas

1. ¿Cuántas fuentes *independientes* cuentan lo mismo que la pieza analizada en el periodo, y desde cuántos países?
2. De las que lo cuentan, ¿cuántas lo apoyan, cuántas lo contradicen y cuántas solo hablan del tema?
3. ¿Cómo se reparte en el tiempo: con precursor y crecimiento escalonado, o de golpe?
4. ¿Qué parte de esas respuestas depende de cómo se busca y no de lo que hay publicado?

4.2 Corpus y fuentes

El corpus es la prensa digital analizada por TrawlingWeb, consultada por su API con sintaxis de palabras clave. TrawlingWeb opera al amparo de la excepción de minería de textos y datos (TDM) del artículo 4 de la Directiva (UE) 2019/790: analiza fuentes públicas y cita cada pieza con su enlace, pero no reproduce ni redistribuye el contenido editorial original. Cada pieza trae título, texto, dominio, autor, idioma, país del medio, fecha de publicación y audiencia estimada del dominio. El análisis usa un periodo de 7 días por defecto (14 como máximo) alrededor de la fecha de la pieza, o hacia atrás desde hoy si solo hay titular. Cada análisis hace como mucho 12 consultas, a una por segundo.

La entrada es un titular con entradilla o un enlace. La propia pieza analizada nunca cuenta como fuente de sí misma.

4.3 Procedimiento

1. **Claves.** Un modelo local de reconocimiento de entidades obtiene personas, organizaciones y lugares, con prioridad para los del titular; unas reglas añaden cifras y fechas.
2. **Búsqueda.** Una consulta de contexto (la entidad más específica con las demás claves) y tres consultas estrictas de dos palabras. Si una consulta se trunca, el periodo se parte en hasta cuatro ventanas para no sesgar hacia las piezas más antiguas.
3. **Coincidencias.** Un modelo local de vectores semánticos multilingües compara cada candidata con el titular y la entradilla. Se aceptan las de similitud ≥ 0,87 que además contienen la entidad ancla.

4. **Independencia.** Las coincidencias se agrupan en fuentes: mismo dominio raíz, misma firma de agencia, mismo autor o texto casi idéntico (≥ 0,85). Cien copias de un teletipo cuentan como una fuente.
5. **Apoyo o contradicción.** Sobre una muestra de 25 a 50 piezas, repartida por día y por fuente y con los desmentidos explícitos primero, un modelo local de inferencia textual decide si cada pieza apoya, contradice o es neutral. Evalúa el título y las tres primeras frases por separado, para que un cuerpo largo no diluya una negación.
6. **Reglas y revisión.** Unas reglas deterministas cubren lo que el modelo no ve: un desmentido explícito («niega que», «desmiente») contradice; la cautela («podría», «presuntamente») y la atribución («según», «denuncia») no apoyan. Las piezas dudosas, las decididas por una regla de parche y las contradicciones (salvo el desmentido explícito) las revisa un modelo de lenguaje que solo clasifica y deja su motivo; no redacta ni aporta fuentes.
7. **Puntuación.** Se calculan las nueve señales, se ponderan y se aplican los recortes.

4.4 Señales y pesos

Señal	Peso	Qué mide
Fuentes independientes	0,20	Número de fuentes, saturado en 5
Apoyo frente a contradicción	0,20	(apoyan − contradicen) / evaluadas, llevado a 0-1
Diversidad de países	0,10	Países distintos entre las fuentes, saturado en 4
Claves en el titular	0,10	Fracción de claves que aparecen en el título de cada coincidencia
Audiencia	0,10	Visitantes únicos del dominio en escala logarítmica
Calidad de la fuente	0,10	Nivel del medio; hoy todos «sin clasificar» (0,2)
Distribución temporal	0,10	Cobertura en al menos 3 tramos del periodo con precursor
Similitud	0,05	Similitud media de las coincidencias aceptadas
Primera aparición	0,05	Si circulaba antes de la pieza analizada

El índice es:

$$\text{índice} = 100 \cdot \sum_{i=1}^{9} w_i \, s_i \,, \qquad \sum_i w_i = 1 \,, \quad s_i \in [0,1]$$

Los pesos son **provisionales**: salen de criterio experto y no están calibrados con datos. Para que una noticia regional cierta no quede penalizada por no tener eco internacional, las saturaciones de fuentes, países y audiencia dependen del alcance esperado (regional, nacional o internacional), que se detecta en el titular.

4.5 Recortes

Condición	El índice no pasa de
Menos de 2 fuentes independientes	25
Ninguna pieza evaluada apoya	25
Contradicen más piezas de las que apoyan	40
Apoyan menos del 25 % de las evaluadas	40

Los recortes son reglas fijadas a mano a partir de casos (sección 6), no el resultado de un ajuste estadístico. Separan «muy difundido» de «corroborado»: un tema caliente con muchas fuentes que no confirman el hecho no puede puntuar alto.

4.6 Salida

Cada análisis devuelve el índice, la etiqueta del tramo (0-25 «sin corroboración suficiente», 25-50 «débil», 50-75 «moderada», 75-100 «alta»), cada señal con su peso, los recortes aplicados, las consultas hechas, cada pieza con su veredicto y su motivo, y los límites del propio análisis. Las etiquetas son de confianza, nunca de veracidad.

5. De la propuesta de 2023 al sistema

De las ideas de 2023, la corroboración y «una IA clasifica, otra revisa» están hechas y medidas; la zona 0 está a medias y la autoridad del emisor sigue sin hacer. Varias piezas del prototipo se sustituyeron porque, al medirlas, no aguantaban.

Idea de 2023	Estado en 2026	Qué cambió y por qué
Buscar quién más cuenta lo mismo en un corpus amplio	Hecho	Búsqueda en la prensa digital analizada con consultas de contexto y estrictas
Contar resultados (umbral de 25 por idioma)	Sustituido	Se cuentan fuentes independientes: el recuento bruto lo infla la sindicación de agencias

Idea de 2023	Estado en 2026	Qué cambió y por qué
Coincidencia por porcentaje de palabras (35 %)	Sustituido	Similitud semántica con umbral medido (0,87); con la léxica, de 124 piezas de la misma noticia contada con otras palabras se aceptaban 2
Una IA clasifica, otra cuestiona (redes adversarias)	Hecho, sin redes adversarias	Inferencia textual local que decide lo claro y un modelo de lenguaje que revisa solo lo dudoso
Zona 0: dónde y cuándo se contó primero	A medias	Dentro del periodo, sí (señal de primera aparición). Hacia atrás, experimental y apagada: falta un caso positivo que la valide
Autoridad del emisor: historial del medio y del autor	No hecho	Pendiente de una lista de calidad de medios; hoy la audiencia hace de sustituto
Traducir el titular a 12 idiomas	No hecho	Las consultas en otros idiomas usan solo entidades; los vectores semánticos cruzan idiomas sin traducir
Puntuación en estrellas con etiqueta «mayormente falsa»	Descartado	Un veredicto de veracidad promete más de lo que se mide
Clasificar el tipo de engaño (Wardle)	Fuera de alcance	Describe el engaño, no la corroboración; podría ser una etiqueta aparte
Coherencia texto-imagen-emoción	Fuera de alcance	El sistema solo trata texto

La lección general es que las ideas sobrevivieron y las implementaciones ingenuas no. Contar resultados, emparejar palabras y emitir un veredicto parecían razonables en una propuesta; con casos reales, cada una fallaba de una forma concreta y medible.

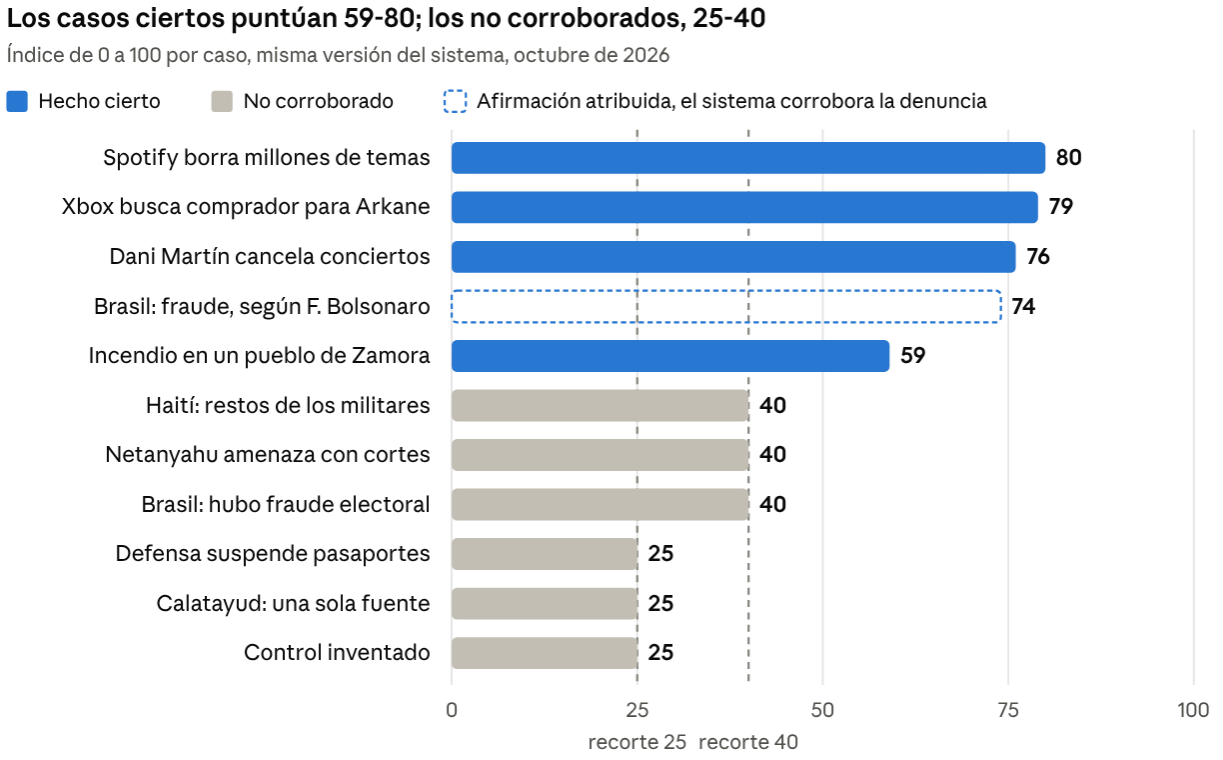
6. Evaluación preliminar

En diez casos medidos con la misma versión, los cuatro hechos ciertos puntuaron entre 59 y 80 y los seis no corroborados entre 25 y 40, sin solaparse. Es una señal de que el índice ordena en la dirección esperada, no una medida de acierto: los casos son pocos y se

eligieron a mano.

6.1 Casos reales

Los casos se eligieron para cubrir tipos distintos: hechos muy difundidos, última hora, noticias regionales, bulos ya desmentidos por verificadores, una afirmación de una sola fuente y un control inventado. Cada uno se analizó con el periodo de su publicación y como mucho 14 consultas.



Registro de mediciones de GeriAI · 11 casos analizados el 30 de octubre de 2026 con la misma versión

Los tres casos no corroborados en 40 y los tres en 25 están ahí por un recorte, no por la suma de señales. Después se añadió el alcance esperado y la noticia regional de Zamora, cubierta por los dos medios locales que la contaron, subió de 59 a 75; los casos internacionales no cambiaron.

Tres casos explican cómo se fijaron los recortes:

- **Desmentido (Haití).** «Los restos hallados son de los militares desaparecidos», que el presidente de El Salvador desmintió días después. La primera versión daba 75: la búsqueda solo traía las piezas más antiguas y ningún desmentido entraba. Con búsqueda por ventanas, muestra repartida por días y reglas de desmentido, queda en 40.
- **Fuente única (Calatayud).** Una noticia local publicada por un solo medio daba 54 porque «se apoyaba a sí misma». De ahí el recorte a 25 con menos de dos fuentes.

- **Tema caliente (Brasil).** «Hubo fraude en las elecciones de Brasil»: 13 fuentes y 9 países, pero solo 2 de 20 piezas evaluadas lo apoyaban. Las señales de difusión sumaban 77. De ahí el recorte por fracción de apoyo: en los casos ciertos esa fracción fue de 0,36 a 0,76 y en los no corroborados, de 0 a 0,12; el umbral de 0,25 cae en el hueco.

La excepción documentada es la atribución. «Hubo fraude en Brasil, *según denuncia* Flávio Bolsonaro» da 74: el sistema corrobora que lo denunció, que es cierto y está muy difundido, no que hubiera fraude. La hipótesis debe ser el hecho, sin «según X».

6.2 Clasificación de apoyo o contradicción

Se compararon evaluadores sobre 251 pares reales (título de una pieza frente al titular analizado) de los casos anteriores: 156 de una primera tanda y 95 añadidos después.

Evaluador	Acierto	Falsas contradicciones	Coste de los 251 pares
Inferencia textual local, sin reglas	0,65	—	0
Inferencia textual local con reglas	0,78	7	0
Modelo de lenguaje económico	0,94	4	0,008 USD

Dos resultados importan más que el acierto global. Primero, 20 de las 37 contradicciones de la primera tanda no las ve ningún modelo de inferencia: son desmentidos del tipo «X niega que…», que en lógica estricta no implican lo contrario. Las resuelve una regla léxica, no un modelo mayor. Segundo, el modelo de lenguaje ve los 39 desmentidos sin reglas.

Aviso sobre estas cifras. Las etiquetas de los 251 pares las puso una IA y no las ha revisado todavía una persona; el modelo de lenguaje evaluado «razona» de forma parecida a quien etiquetó, así que su 0,94 está probablemente inflado. Además, la prueba usa solo títulos, mientras que el sistema usa título y tres frases. Sirve para comparar evaluadores entre sí, no como acierto esperado.

6.3 El sistema híbrido en casos reales

En seis casos con la misma muestra, la revisión de las piezas dudosas por el modelo de lenguaje no movió los controles (Spotify, BCE, Dani Martín, Haití y Brasil quedaron igual) y corrigió un caso: Nvidia pasó de 75 a 81 porque la regla de cautela («presunta agente») dejaba como neutrales 15 piezas que sí confirmaban la investigación. Revisó entre 2 y 15 piezas por análisis, con un coste de 0,0003 a 0,0016 USD y de 0 a 3 segundos más.

6.4 Lo que enseñó el uso real

Con el sistema abierto al público, los análisis reales destaparon fallos que los casos elegidos no mostraban. Tres ejemplos: el último día del periodo quedaba fuera de la búsqueda, así que una noticia del día no encontraba sus ecos (de 0 a 53 al corregirlo); con entrada por enlace, las entidades del cuerpo arrastraban la búsqueda a otro tema; y piezas anteriores al hecho («pendiente de una votación») se leían como contradicciones de que el hecho ocurriera. Cada fallo se corrigió y se midió con un caso y dos controles.

7. Límites y amenazas a la validez

Lo que este documento permite afirmar es poco y concreto: el índice ordena en la dirección esperada en unos diez casos, y cada regla resuelve el fallo que la motivó sin romper los controles. No permite afirmar una tasa de acierto ni que el índice generalice.

- **Muestra pequeña y elegida.** Unos diez casos de calibración y una veintena de análisis reales, elegidos por tipo y no al azar. Los mismos casos sirvieron para fijar reglas y para comprobarlas, lo que favorece el ajuste a esos casos.
- **Pesos sin calibrar.** Los nueve pesos salen de criterio experto. Otro reparto razonable podría cambiar el orden de casos cercanos.
- **Etiquetas sin revisión humana.** Ver el aviso de la sección 6.2.
- **Sin referencia externa.** No se ha comparado el índice con el juicio de verificadores profesionales ni con conjuntos públicos de verificación.
- **Dependencia del corpus.** Si una pieza no está en el corpus analizado o la búsqueda no la encuentra, el sistema no puede evaluarla: un 0 puede significar «no hay nada» o «no lo encontró». Un hecho contado con palabras muy distintas puede perder piezas.
- **Fechas ruidosas.** La fecha de publicación que dan los medios no siempre es fiable, y la fecha de la pieza analizada no es la fecha del hecho; eso limita las señales temporales.
- **Lo que no mide.** No detecta sátira, no analiza imagen ni vídeo, no mide la autoridad del autor y no sabe si una afirmación es una opinión o una tesis sin hecho verificable.
- **Corroboración no es verdad.** Un bulo copiado por muchas fuentes independientes que no lo desmienten puntuaría alto. El índice mide cuánto se corrobora, y así se presenta siempre.

8. Qué queda abierto

El siguiente paso necesario es un conjunto de evaluación con etiquetas humanas; sin él, ninguna de las preguntas siguientes se puede cerrar.

1. **Evaluación con referencia externa.** Un conjunto de al menos 300 pares con etiquetas revisadas por dos personas, y una muestra de afirmaciones ya verificadas por verificadores profesionales para comparar el índice con su juicio.
2. **Calibración de pesos.** Ajustar los nueve pesos con ese conjunto y publicar cuánto cambia el orden de los casos.
3. **Zona 0 hacia atrás.** Buscar la primera aparición real fuera del periodo analizado. Está construida y apagada: le falta un caso positivo que la valide. Una vía es usar como fecha la de la primera observación propia en el archivo continuo, más fiable que la fecha que declara cada medio, aunque depende de cuántos meses cubra ese archivo.
4. **Autoridad del emisor.** Saber si el medio o el autor ya trataba el tema antes y con qué consistencia, una vez exista una lista de calidad de medios.
5. **Fecha del hecho.** Estimar cuándo ocurrió algo, y no solo cuándo se publicó la pieza, para separar lo anterior al hecho de lo que lo contradice.
6. **Opiniones y tesis.** Detectar titulares sin hecho verificable y avisar en lugar de puntuar. En la misma línea, añadir el tipo de engaño según Wardle como etiqueta aparte, sin que mueva el índice.
7. **El efecto en las personas.** Cómo cambia lo que creemos y compartimos cuando una información llega con su índice y sus pruebas. Hay estudios sobre cómo evaluamos la credibilidad, pero ningún trabajo propio todavía.

Cómo citar

Trabazos Clavería, Ó. (2026). *Medir la corroboración, no la veracidad: un índice de confianza con pruebas para la prensa digital* (working paper, versión 1.0, sin revisión por pares). https://oscartrabazos.com/papers/trabazos_2026_corroboracion_es.pdf

Referencias

- Ahmed, H. (2017). *Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity*. Tesis de máster, Universidad de Victoria.
- Bowman, S. R., Angeli, G., Potts, C. y Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of EMNLP 2015*.
- Shu, K., Sliva, A., Wang, S., Tang, J. y Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations*, 19(1).
- Sirajudeen, S. M., Azmi, N. F. A. y Abubakar, A. I. (2017). Online fake news detection algorithm. *Journal of Theoretical and Applied Information Technology*, 95(17).
- Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S. y de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. arXiv:1704.07506.

- Thorne, J., Vlachos, A., Christodoulopoulos, C. y Mittal, A. (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. *Proceedings of NAACL-HLT 2018*.
- Trabazos, O., Suárez, S., Bori, R. y Flo, O. (2014). Aplicación de tecnologías de procesamiento de lenguaje natural y tecnología semántica en Brand Rain y Anpro21. *Procesamiento del Lenguaje Natural*, 53, 201-204.
- Wardle, C. y Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Consejo de Europa. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c
- Wardle, C. (16 de febrero de 2017). Fake news. It's complicated. First Draft. Recoge las «cuatro P» que propuso Eliot Higgins y las amplía. https://firstdraftnews.org/articles/fake-news-complicated/

Lecturas de introducción

- Wikipedia: Eliot Higgins, fundador de Bellingcat.
- Wikipedia: Fake news, que resume las siete categorías de Wardle.

Nota sobre las fuentes propias

La trayectoria (sección 3) se apoya en el blog del autor (2009-2026), en el artículo de la SEPLN de 2014, en documentos de la sociedad NAD (2018-2020), en dos borradores inéditos de mayo de 2018 y en material no publicado de 2023 (el abstract de la técnica del impostor y el esquema del prototipo). El método y la evaluación (secciones 4 a 6) salen del registro de mediciones de GeriAI, donde cada caso y cada cambio están anotados con fecha, versión y resultado.