Mesurar la corroboració, no la veracitat: un índex de confiança amb proves per a la premsa digital

Óscar Trabazos Clavería · Working paper, sense revisió per parells · Versió 1.0, 2 de novembre de 2026

Resum

Aquest document proposa mesurar quanta corroboració té una informació a la premsa digital, en lloc de declarar-la vertadera o falsa. Descriu GeriAI, un sistema que, a partir d'un titular o d'un enllaç, cerca peces que expliquen el mateix, agrupa les còpies d'una mateixa font, decideix si cada peça hi dóna suport, la contradiu o no en diu res, i combina nou senyals en un índex de 0 a 100 que sempre arriba amb les seves proves.

El treball recull una línia que va començar el 2007 mesurant el to i la reputació en allò que es publica, que el 2017 es va preguntar «com es pot detectar una notícia falsa?» i que el 2026 es va reformular com a «quanta confiança mereix i amb quines proves?». Presenta el mètode, la traducció d'una proposta sense mesurar del 2023 (la «tècnica de l'impostor») a un sistema mesurat, i una avaluació preliminar: una desena de casos reals i una prova de classificació sobre 251 parells de titulars.

L'avaluació és preliminar i ho diu: els pesos són provisionals i no estan calibrats, els casos són exemples i no una mostra, i una part de les etiquetes de la prova les va posar una IA sense revisió humana encara. El document separa el que s'ha fet i mesurat del que es va proposar i del que queda obert.

Paraules clau: desinformació, corroboració, fonts independents, inferència textual, índex de confiança, premsa digital.

Sobre l'autor

Óscar Trabazos Clavería construeix des del 2007 sistemes que llegeixen el que es publica a la premsa i a internet i mesuren alguna cosa més que paraules: primer el to i la reputació de les marques, després la ironia i l'emoció, i des del 2017 la desinformació.

- **Anpro21 i Brand Rain.** A Anpro21 va llançar el 2009 una mesura del to de les marques a la premsa i a internet, i el 2010 Brand Rain, una eina de mesura de la reputació. La seva base tècnica es va publicar a la revista *Procesamiento del Lenguaje Natural* de la SEPLN (2014).

- **TrawlingWeb.** N'és el fundador i conseller delegat. TrawlingWeb analitza premsa digital i xarxes socials de desenes de països, i GeriAI n'és la capa d'anàlisi amb IA. El sistema d'aquest document treballa sobre aquest corpus.
- **Recerca amb universitats.** És coautor, amb Pilar Sánchez-García (Universidad de Valladolid) i João Canavilhas (Universidade da Beira Interior), d'un estudi sobre com la premsa d'Espanya i Portugal explica la IA, ara en avaluació. El setembre del 2026 va fer la conferència inaugural del I Congrés Internacional IBERPERIA (Covilhã).
- **Escriptura.** Publica al seu blog des del 2006; la sèrie sobre desinformació va del 2017 al 2026.

Aquest working paper és la primera vegada que la línia sobre desinformació es presenta amb mètode i mesures. El que hi havia abans eren textos de divulgació, propostes i prototips, i així es tracten aquí.

1. Introducció

La pregunta d'aquest treball és quanta confiança mereix una informació publicada, mesurada per quant la corroboren fonts independents de la premsa digital en un període determinat. No pregunta si és vertadera.

Els enganys no són nous. El que és nou és la velocitat: qualsevol publica, ningú no ha de contrastar abans, i el que encaixa amb el que el públic vol sentir es multiplica tot sol. La resposta clàssica del periodisme continua valent: contrastar amb fonts independents abans de donar res per bo. El que canvia és l'escala a la qual cal fer-ho.

Un veredicte binari falla en els dos sentits. Un fet cert però recent o local pot tenir poca cobertura; un bulo molt copiat en pot tenir molta. Per això el sistema no diu «vertader» ni «fals»: dóna un índex de 0 a 100, l'etiqueta del seu tram (de «sense corroboració suficient» a «corroboració alta») i totes les peces que el sostenen o el contradiuen, perquè jutgi qui llegeix.

Aquest document aporta tres coses:

1. Un mètode explícit per mesurar la corroboració: com es cerquen les peces, com es compten les fonts independents, com es decideix si hi donen suport o la contradiuen i com es combinen els senyals.
2. El recorregut d'una proposta metodològica del 2023, que no tenia xifres, fins a un sistema en producció amb cada canvi mesurat en casos reals.
3. Una avaluació preliminar amb els seus límits a la vista.

No aporta una avaluació sistemàtica: no hi ha un conjunt de prova ampli amb etiquetes revisades per persones ni una comparació amb verificadors professionals. La secció 7 detalla què es pot concloure i què no.

2. Antecedents i marc

El treball es recolza en quatre idees de la literatura i en una regla de l'ofici. No és una revisió sistemàtica: cita el que n'ha orientat el disseny.

- **Tipus de desordre informatiu.** Wardle i Derakhshan (2017) distingeixen informació errònia, desinformació i malinformació, i set formes: sàtira, connexió falsa, contingut enganyós, context fals, contingut impostor, contingut manipulat i contingut inventat. Aquest marc descriu el *tipus* d'engany; l'índex d'aquest treball no el classifica, mesura una altra cosa (la corroboració).
- **La motivació de l'emissor.** Les «quatre P» que Eliot Higgins atribueix a la desinformació (passió, política, propaganda i pagament), recollides per Wardle (2017), expliquen per què es fabrica. Aquí no es mesuren: queden com a context.
- **El text sol no n'hi ha prou.** Shu et al. (2017) revisen la detecció automàtica a les xarxes i conclouen que el contingut s'ha de completar amb el context social: fonts, xarxa i difusió. Tacchini et al. (2017) classifiquen publicacions per qui hi interactua, no pel que diuen. Ahmed (2017) mostra l'abast i el límit dels trets lèxics (n-grames). Aquest treball en pren la conclusió comuna: el que delata una informació és sobretot fora del seu text.
- **Verificació d'afirmacions amb proves.** La verificació automàtica d'afirmacions (per exemple, FEVER, Thorne et al., 2018) recupera textos i decideix si donen suport a una afirmació o la refuten, amb la inferència textual (Bowman et al., 2015) com a peça central. GeriAI fa servir aquest esquema, però sobre premsa del mateix període i per mesurar corroboració, no per emetre un veredicte.
- **La regla més antiga del periodisme.** Contrastar amb fonts independents abans de publicar. L'índex la converteix en nombres: quantes fonts diferents, de quants països, quantes hi donen suport i quantes la contradiuen.

La diferència amb la major part de la literatura de detecció és l'objecte de la mesura. No s'entrena un classificador de «fals/vertader»; es compta i es qualifica la corroboració disponible, i s'ensenya.

3. Trajectòria de la línia (2007-2026)

La pregunta va canviar tres vegades: d'entendre el to i la reputació (2007-2014) a detectar el que és fals (2017-2023) i d'aquí a mesurar la confiança (2025-2026). En desinformació, fins al 2026 tot van ser textos de divulgació, propostes i prototips; les primeres mesures són les d'aquest document.

Any	Fita	Què aporta a aquest treball	Tipus de font
2007-2009	A Anpro21, primer intent d'un algorisme que entengués el sentiment del que es publica; el 2009, una mesura del «to» de cada marca a la premsa i a internet	La pregunta de fons: mesurar a escala el que es publica. I el primer mur, el sarcasme, que una màquina d'aleshores no percebia	Blog (2009)
2010-2011	Brand Rain: reputació i percepció de marca; estudi de les eleccions generals del 20N amb uns 15.000 missatges en tres setmanes	Mesurar alhora premsa, blogs i xarxes, i aplicar-ho a la política	Blog (2011)
2012	Projecte de detecció d'amenaces, segons l'autor per encàrrec del Ministeri de l'Interior a través d'una empresa sòcia que desenvolupava programari de defensa	Segons l'autor, l'origen de la idea de «detectar l'impostor»	Relat de l'autor, sense document
2013	Ironia, sarcasme i to en la mesura d'intangibles	La ironia com l'error més visible de qualsevol mesura automàtica d'opinió	Blog (2013)
2014	Article a *Procesamiento del Lenguaje Natural* núm. 53 (SEPLN) sobre PLN i tecnologia semàntica a Brand Rain i Anpro21	La base tècnica: escolta activa, entitats, relacions i xarxes complexes sobre premsa. No tracta de desinformació	Publicació revisada
2014	Projecte de vídeo-reputació amb Ibermática, Verbio, CSA i el grup BISITE (Universidad de Salamanca), amb el suport del Pla Avança	Si el text i la imatge no quadren, cal mirar dues vegades: la idea torna el 2024 amb el vídeo sintètic	Blog (2014)

Any	Fita	Què aporta a aquest treball	Tipus de font
2017	«Las fake news en la política: un desafío para la democracia» (blog)	Comença la línia sobre desinformació: el problema és la difusió, no l'engany en si	Divulgació
2017-2020	Jimcrick, aplicació contra l'assetjament escolar de la societat NAD	Detectar amenaces en les converses de menors; segons l'autor, amb la mateixa idea de l'impostor	Documents de NAD (2018-2020)
2018	Sèrie «Políticos, fake news y big data» (2 parts, inèdita fins ara; signada aleshores amb pseudònim)	Presenta l'algorisme de Sirajudeen et al. (2017): comprovar la font i cercar si el títol i el tema apareixen en altres llocs. És la primera vegada que la corroboració externa apareix a la línia, com a mètode d'altri	Esborrany propi sobre treball d'altri
2023	Sèrie «Detectar fake news» I-III: titulars semàntics, classificació en dues etapes, Wardle i Higgins	El marc teòric i la tècnica de l'impostor (apartat 3.1)	Divulgació + abstract inèdit
2023	Prototip conversacional sobre el corpus de TrawlingWeb, provat amb l'acomiadament de Sam Altman	Cercava el titular en 12 idiomes i puntuava amb estrelles. Sense codi de motor ni mesures	Prototip + vídeos de demostració
2024	«Sora: cambiando las reglas del juego de la desinformación»	La coherència entre text, imatge i emoció com a senyal davant del vídeo sintètic	Divulgació
2025	«Seguretat temàtica» per a qui redacta	Tres eixos: abast, credibilitat («qui més parla del tema?») i autoritat de l'emissor. El segon eix és l'índex d'aquest treball	Divulgació

Any	Fita	Què aporta a aquest treball	Tipus de font
2026	Sèrie sobre la desinformació com a risc i la verificació amb models generatius	El gir: «no cerquem la veritat»; patrons de propagació (velocitat davant de volum, homogeneïtat, historial de la font)	Divulgació
2026	GeriAI: índex de corroboració en producció, amb registre de mesures	El sistema i l'avaluació de les seccions 4 a 6	Sistema + registre de mesures

Dos avisos sobre aquesta trajectòria. Primer, alguns textos de divulgació del 2023 prometien més del que existia (validar «en tres minuts» si una cosa era falsa, xarxes adversarials, cadena de blocs); res d'això no es va construir i aquest document no ho recull com a resultat. Segon, l'abstract de la tècnica de l'impostor (2023) en situa l'origen en un projecte de detecció d'amenaces del 2012 (segons l'autor, un encàrrec del Ministeri de l'Interior a través d'una empresa sòcia que desenvolupava programari de defensa; l'abstract l'atribuïa per error a l'OTAN) i en el seu ús posterior a Jimcrick, una aplicació contra l'assetjament escolar. Que Jimcrick va existir i era una aplicació contra l'assetjament consta en documents del 2018 al 2020; que fes servir aquesta tècnica, i el projecte del 2012, no tenen document. Per això figuren a la taula com a relat de l'autor i no sostenen cap conclusió d'aquest treball.

3.1 La tècnica de l'impostor (2023)

La tècnica de l'impostor proposa jutjar una informació per qui l'emet i per on va començar, no només pel que diu. Es va formular en un abstract de novembre del 2023 que no es va arribar a publicar i es va divulgar a la tercera part de la sèrie «Detectar fake news».

La idea de partida és interrogar tots els qui participen en una conversa (usuaris, periodistes, mitjans, autoritats) per descobrir qui no és el que diu ser: l'impostor. Segons l'autor, va néixer en el projecte de detecció d'amenaces del 2012 i es va reutilitzar contra l'assetjament escolar. L'abstract l'aplica a la desinformació amb tres components:

1. **Zona 0.** Resseguir enrere fins a l'instant, el lloc i la font que van tractar el tema per primera vegada. Qui ho va explicar primer i com es va estendre a partir d'allà diuen més que el text de qualsevol còpia.
2. **Autoritat de l'emissor.** Revisar l'historial del mitjà i de l'autor que signa: si ja tractaven aquest tema, amb quina consistència i amb quina reputació en aquest àmbit.

3. **IA discriminativa i IA adversarial.** Una IA classifica i una altra qüestiona el que conclou la primera, per reduir els falsos positius.

El marc teòric combina Wardle, per saber *quin tipus* d'engany hi ha, i Higgins, per preguntar-se *per què* es difon (passió, política, propaganda o pagament). L'autoritat de l'emissor és la resposta pràctica a una de les formes de Wardle, el contingut impostor, que dóna nom a la tècnica.

El seu límit és que era una proposta sense mesurar. L'abstract afirma que la tècnica «ha demostrat ser una eina eficac», però no dóna cap xifra; la secció de resultats enumera línies futures i la de referències va quedar per escriure. El mateix text reconeix que no s'havia comparat amb altres tècniques. Aquest treball pren les tres idees i les sotmet a mesura (secció 5).

3.2 Quins tipus de Wardle pot veure la corroboració

La corroboració veu bé el contingut inventat i molt poc de la resta. Mesurar fins a quin punt es confirma un fet no diu, per si sol, de quina manera enganya una peça.

Tipus (Wardle i Derakhshan, 2017)	Què pot veure l'índex	Per què
Contingut inventat	Sí	Sense fonts independents que l'expliquin, l'índex no passa de 25 (control inventat i font única, secció 6)
Contingut enganyós	En part	Les regles de xifra diferent, atribució i cautela eviten comptar com a suport el que no ho és; no etiqueten l'engany
Connexió falsa	En part	En cada coincidència s'avaluen el títol i el cos per separat; no es comprova si el titular analitzat resumeix el seu propi cos
Contingut impostor	Poc	Un mitjà suplantat que publica sol queda en «una sola font»; detectar-lo de debò exigeix l'autoritat de l'emissor, que no està feta
Context fals	No	Si el fet és cert, es corrobora encara que es presenti en un context fals
Contingut manipulat	No	El sistema només tracta text; no analitza imatges ni vídeo
Sàtira o paròdia	No	Pot sortir sense corroboració, però el sistema no sap que és sàtira

Per això l'índex no substitueix la classificació de Wardle, sinó que la complementa. El tipus d'engany es podria afegir com una etiqueta a part, sense barrejar-la amb el nombre (secció 8).

4. Mètode

L'índex és una suma ponderada de nou senyals de corroboració, retallada per quatre regles fixes i acompanyada de totes les peces que la sostenen. Cada llindar i cada regla es van fixar mesurant casos reals; els pesos, no.

4.1 Preguntes

1. Quantes fonts *independents* expliquen el mateix que la peça analitzada en el període, i des de quants països?
2. De les que ho expliquen, quantes hi donen suport, quantes la contradiuen i quantes només parlen del tema?

3. Com es reparteix en el temps: amb un precursor i un creixement esglaonat, o de cop?
4. Quina part d'aquestes respostes depèn de com es cerca i no del que s'ha publicat?

4.2 Corpus i fonts

El corpus és la premsa digital analitzada per TrawlingWeb, consultada per la seva API amb sintaxi de paraules clau. TrawlingWeb opera a l'empara de l'excepció de mineria de textos i dades (TDM) de l'article 4 de la Directiva (UE) 2019/790: analitza fonts públiques i cita cada peça amb el seu enllaç, però no reprodueix ni redistribueix el contingut editorial original. Cada peça porta títol, text, domini, autor, idioma, país del mitjà, data de publicació i audiència estimada del domini. L'anàlisi fa servir un període de 7 dies per defecte (14 com a màxim) al voltant de la data de la peça, o enrere des d'avui si només hi ha titular. Cada anàlisi fa com a màxim 12 consultes, a una per segon.

L'entrada és un titular amb entradeta o un enllaç. La mateixa peça analitzada mai no compta com a font d'ella mateixa.

4.3 Procediment

1. **Claus.** Un model local de reconeixement d'entitats obté persones, organitzacions i llocs, amb prioritat per als del titular; unes regles hi afegeixen xifres i dates.
2. **Cerca.** Una consulta de context (l'entitat més específica amb les altres claus) i tres consultes estrictes de dues paraules. Si una consulta es trunca, el període es parteix en fins a quatre finestres per no esbiaixar cap a les peces més antigues.
3. **Coincidències.** Un model local de vectors semàntics multilingües compara cada candidata amb el titular i l'entradeta. S'accepten les de similitud ≥ 0,87 que, a més, contenen l'entitat àncora.
4. **Independència.** Les coincidències s'agrupen en fonts: mateix domini arrel, mateixa signatura d'agència, mateix autor o text gairebé idèntic (≥ 0,85). Cent còpies d'un teletip compten com una sola font.
5. **Suport o contradicció.** Sobre una mostra de 25 a 50 peces, repartida per dia i per font i amb els desmentiments explícits primer, un model local d'inferència textual decideix si cada peça hi dóna suport, la contradiu o és neutral. Avalua el títol i les tres primeres frases per separat, perquè un cos llarg no dilueixi una negació.
6. **Regles i revisió.** Unes regles deterministes cobreixen el que el model no veu: un desmentiment explícit («nega que», «desmenteix») contradiu; la cautela («podria», «presumptament») i l'atribució («segons», «denuncia») no hi donen suport. Les peces dubtoses, les decidides per una regla de pedaç i les contradiccions (excepte el desmentiment explícit) les revisa un model de llenguatge que només classifica i deixa el seu motiu; no redacta ni aporta fonts.
7. **Puntuació.** Es calculen els nou senyals, es ponderen i s'apliquen els retalls.

4.4 Senyals i pesos

Senyal	Pes	Què mesura
Fonts independents	0,20	Nombre de fonts, saturat en 5
Suport davant de contradicció	0,20	(suport − contradicció) / avaluades, portat a 0-1
Diversitat de països	0,10	Països diferents entre les fonts, saturat en 4
Claus al titular	0,10	Fracció de claus que apareixen al títol de cada coincidència
Audiència	0,10	Visitants únics del domini en escala logarítmica
Qualitat de la font	0,10	Nivell del mitjà; avui tots «sense classificar» (0,2)
Distribució temporal	0,10	Cobertura en almenys 3 trams del període amb precursor
Similitud	0,05	Similitud mitjana de les coincidències acceptades
Primera aparició	0,05	Si circulava abans de la peça analitzada

L'índex és:

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

Els pesos són **provisionals**: surten de criteri expert i no estan calibrats amb dades. Perquè una notícia regional certa no quedi penalitzada per no tenir resson internacional, les saturacions de fonts, països i audiència depenen de l'abast esperat (regional, nacional o internacional), que es detecta al titular.

4.5 Retalls

Condició	L'índex no passa de
Menys de 2 fonts independents	25
Cap peça avaluada no hi dóna suport	25
La contradiuen més peces de les que hi donen suport	40
Hi donen suport menys del 25 % de les avaluades	40

Els retalls són regles fixades a mà a partir de casos (secció 6), no el resultat d'un ajust estadístic. Separen «molt difos» de «corroborat»: un tema calent amb moltes fonts que no confirmen el fet no pot puntuar alt.

4.6 Sortida

Cada anàlisi retorna l'índex, l'etiqueta del tram (0-25 «sense corroboració suficient», 25-50 «feble», 50-75 «moderada», 75-100 «alta»), cada senyal amb el seu pes, els retalls aplicats, les consultes fetes, cada peça amb el seu veredicte i el seu motiu, i els límits de la mateixa anàlisi. Les etiquetes són de confiança, mai de veracitat.

5. De la proposta del 2023 al sistema

De les idees del 2023, la corroboració i «una IA classifica, una altra revisa» estan fetes i mesurades; la zona 0 està a mitges i l'autoritat de l'emissor encara no està feta. Diverses peces del prototip es van substituir perquè, en mesurar-les, no aguantaven.

Idea del 2023	Estat el 2026	Què va canviar i per què
Cercar qui més explica el mateix en un corpus ampli	Fet	Cerca a la premsa digital analitzada amb consultes de context i estrictes
Comptar resultats (llindar de 25 per idioma)	Substituït	Es compten fonts independents: la sindicació d'agències infla el recompte brut
Coincidència per percentatge de paraules (35 %)	Substituït	Similitud semàntica amb un llindar mesurat (0,87); amb la lèxica, de 124 peces que explicaven la mateixa notícia amb altres paraules se n'acceptaven 2
Una IA classifica, una altra qüestiona (xarxes adversarials)	Fet, sense xarxes adversarials	Inferència textual local que decideix el que és clar i un model de llenguatge que revisa només el que és dubtós
Zona 0: on i quan es va explicar primer	A mitges	Dins del període, sí (senyal de primera aparició). Enrere, experimental i apagada: hi falta un cas positiu que la validi

Idea del 2023	Estat el 2026	Què va canviar i per què
Autoritat de l'emissor: historial del mitjà i de l'autor	No fet	Pendent d'una llista de qualitat de mitjans; avui l'audiència en fa de substitut
Traduir el titular a 12 idiomes	No fet	Les consultes en altres idiomes fan servir només entitats; els vectors semàntics creuen idiomes sense traduir
Puntuació amb estrelles i etiqueta «majoritàriament fals»	Descartat	Un veredicte de veracitat promet més del que es mesura
Classificar el tipus d'engany (Wardle)	Fora d'abast	Descriu l'engany, no la corroboració; podria ser una etiqueta a part
Coherència text-imatge-emoció	Fora d'abast	El sistema només tracta text

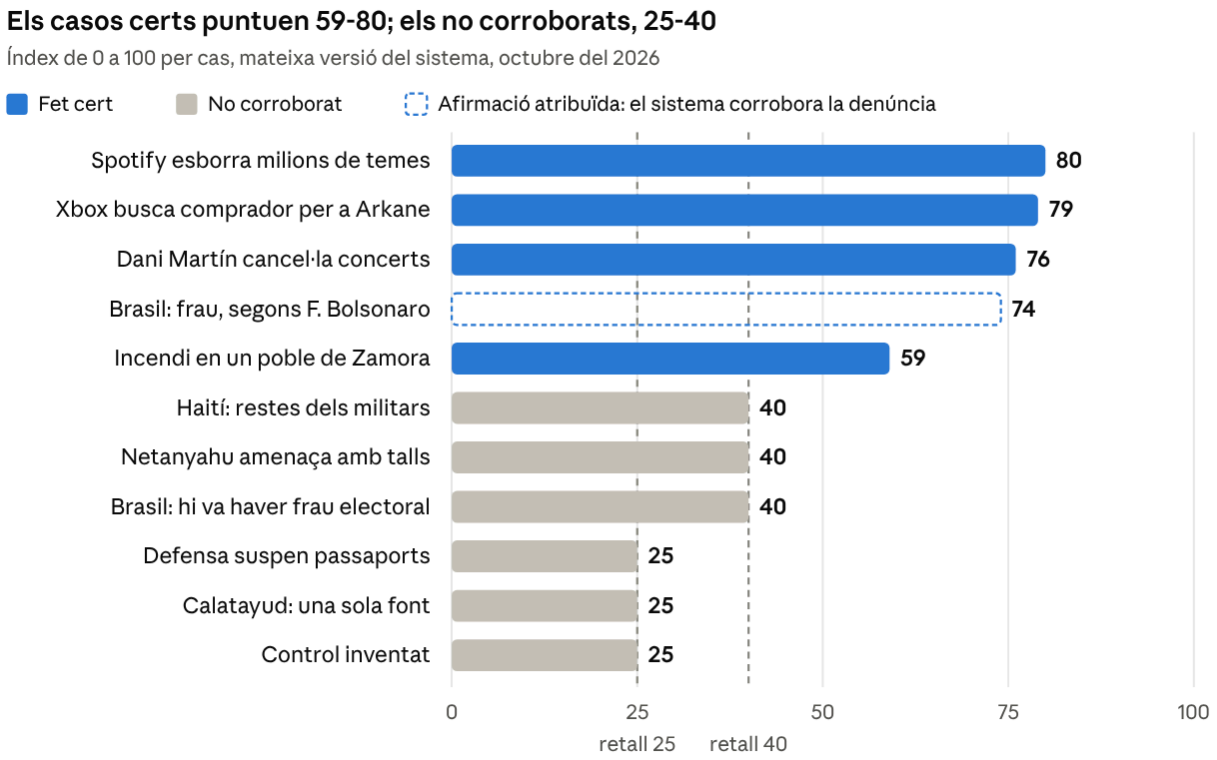
La lliçó general és que les idees van sobreviure i les implementacions ingènues no. Comptar resultats, aparellar paraules i emetre un veredicte semblaven raonables en una proposta; amb casos reals, cadascuna fallava d'una manera concreta i mesurable.

6. Avaluació preliminar

En deu casos mesurats amb la mateixa versió, els quatre fets certs van puntuar entre 59 i 80 i els sis no corroborats entre 25 i 40, sense encavalcar-se. És un senyal que l'índex ordena en la direcció esperada, no una mesura d'encert: els casos són pocs i es van triar a mà.

6.1 Casos reals

Els casos es van triar per cobrir tipus diferents: fets molt difosos, última hora, notícies regionals, bulos ja desmentits per verificadors, una afirmació d'una sola font i un control inventat. Cadascun es va analitzar amb el període de la seva publicació i com a màxim 14 consultes.



Registre de mesures de GeriAI · 11 casos analitzats el 30 d'octubre del 2026 amb la mateixa versió

Els tres casos no corroborats a 40 i els tres a 25 hi són per un retall, no per la suma de senyals. Després s'hi va afegir l'abast esperat, i la notícia regional de Zamora, coberta pels dos mitjans locals que la van explicar, va pujar de 59 a 75; els casos internacionals no van canviar.

Tres casos expliquen com es van fixar els retalls:

- **Desmentiment (Haití).** «Les restes trobades són dels militars desapareguts», que el president del Salvador va desmentir dies després. La primera versió donava 75: la cerca només portava les peces més antigues i no hi entrava cap desmentiment. Amb cerca per finestres, mostra repartida per dies i regles de desmentiment, queda en 40.
- **Font única (Calatayud).** Una notícia local publicada per un sol mitjà donava 54 perquè «es donava suport a si mateixa». Per això el retall a 25 amb menys de dues fonts.
- **Tema calent (Brasil).** «Hi va haver frau a les eleccions del Brasil»: 13 fonts i 9 països, però només 2 de 20 peces avaluades hi donaven suport. Els senyals de difusió sumaven 77. Per això el retall per fracció de suport: en els casos certs aquesta fracció va ser de 0,36 a 0,76 i en els no corroborats, de 0 a 0,12; el llindar de 0,25 cau en el buit.

L'excepció documentada és l'atribució. «Hi va haver frau al Brasil, *segons denuncia* Flávio Bolsonaro» dóna 74: el sistema corrobora que ho va denunciar, que és cert i està molt difós, no que hi hagués frau. La hipòtesi ha de ser el fet, sense «segons X».

6.2 Classificació de suport o contradicció

Es van comparar avaluadors sobre 251 parells reals (el títol d'una peça davant del titular analitzat) dels casos anteriors: 156 d'una primera tanda i 95 afegits després.

Avaluador	Encert	Falses contradiccions	Cost dels 251 parells
Inferència textual local, sense regles	0,65	—	0
Inferència textual local amb regles	0,78	7	0
Model de llenguatge econòmic	0,94	4	0,008 USD

Dos resultats importen més que l'encert global. Primer, 20 de les 37 contradiccions de la primera tanda no les veu cap model d'inferència: són desmentiments del tipus «X nega que…», que en lògica estricta no impliquen el contrari. Les resol una regla lèxica, no un model més gran. Segon, el model de llenguatge veu els 39 desmentiments sense regles.

Avís sobre aquestes xifres. Les etiquetes dels 251 parells les va posar una IA i encara no les ha revisades cap persona; el model de llenguatge avaluat «raona» de manera semblant a qui les va etiquetar, de manera que el seu 0,94 probablement està inflat. A més, la prova fa servir només títols, mentre que el sistema fa servir títol i tres frases. Serveix per comparar avaluadors entre si, no com a encert esperat.

6.3 El sistema híbrid en casos reals

En sis casos amb la mateixa mostra, la revisió de les peces dubtoses pel model de llenguatge no va moure els controls (Spotify, BCE, Dani Martín, Haití i Brasil van quedar igual) i va corregir un cas: Nvidia va passar de 75 a 81 perquè la regla de cautela («presumpta agent») deixava com a neutrals 15 peces que sí que confirmaven la investigació. Va revisar entre 2 i 15 peces per anàlisi, amb un cost de 0,0003 a 0,0016 USD i de 0 a 3 segons més.

6.4 El que va ensenyar l'ús real

Amb el sistema obert al públic, les anàlisis reals van destapar errors que els casos triats no mostraven. Tres exemples: l'últim dia del període quedava fora de la cerca, de manera que una notícia del dia no trobava els seus ressons (de 0 a 53 en corregir-ho); amb entrada per enllaç, les entitats del cos arrossegaven la cerca cap a un altre tema; i peces anteriors al fet («pendent d'una votació») es llegien com a contradiccions que el fet hagués passat. Cada error es va corregir i es va mesurar amb un cas i dos controls.

7. Límits i amenaces a la validesa

El que aquest document permet afirmar és poc i concret: l'índex ordena en la direcció esperada en una desena de casos, i cada regla resol l'error que la va motivar sense trencar els controls. No permet afirmar una taxa d'encert ni que l'índex generalitzi.

- **Mostra petita i triada.** Una desena de casos de calibratge i una vintena d'anàlisis reals, triats per tipus i no a l'atzar. Els mateixos casos van servir per fixar regles i per comprovar-les, cosa que afavoreix l'ajust a aquests casos.
- **Pesos sense calibrar.** Els nou pesos surten de criteri expert. Un altre repartiment raonable podria canviar l'ordre de casos propers.
- **Etiquetes sense revisió humana.** Vegeu l'avís de la secció 6.2.
- **Sense referència externa.** No s'ha comparat l'índex amb el judici de verificadors professionals ni amb conjunts públics de verificació.
- **Dependència del corpus.** Si una peça no és al corpus analitzat o la cerca no la troba, el sistema no la pot avaluar: un 0 pot voler dir «no hi ha res» o «no ho ha trobat». Un fet explicat amb paraules molt diferents pot perdre peces.
- **Dates sorolloses.** La data de publicació que donen els mitjans no sempre és fiable, i la data de la peça analitzada no és la data del fet; això limita els senyals temporals.
- **El que no mesura.** No detecta sàtira, no analitza imatge ni vídeo, no mesura l'autoritat de l'autor i no sap si una afirmació és una opinió o una tesi sense cap fet verificable.
- **Corroboració no és veritat.** Un bulo copiat per moltes fonts independents que no el desmenteixen puntuaria alt. L'índex mesura quant es corrobora una cosa, i sempre es presenta així.

8. Què queda obert

El pas següent necessari és un conjunt d'avaluació amb etiquetes humanes; sense aquest conjunt, cap de les preguntes següents no es pot tancar.

1. **Avaluació amb referència externa.** Un conjunt d'almenys 300 parells amb etiquetes revisades per dues persones, i una mostra d'afirmacions ja verificades per verificadors professionals per comparar l'índex amb el seu judici.
2. **Calibratge de pesos.** Ajustar els nou pesos amb aquest conjunt i publicar quant canvia l'ordre dels casos.
3. **Zona 0 enrere.** Cercar la primera aparició real fora del període analitzat. Està construïda i apagada: hi falta un cas positiu que la validi. Una via és fer servir com a data la de la primera observació pròpia a l'arxiu continu, més fiable que la data que declara cada mitjà, tot i que depèn de quants mesos cobreixi aquest arxiu.

4. **Autoritat de l'emissor.** Saber si el mitjà o l'autor ja tractava el tema abans i amb quina consistència, un cop hi hagi una llista de qualitat de mitjans.
5. **Data del fet.** Estimar quan va passar una cosa, i no només quan es va publicar la peça, per separar el que és anterior al fet del que el contradiu.
6. **Opinions i tesis.** Detectar titulars sense fet verificable i avisar en lloc de puntuar. En la mateixa línia, afegir el tipus d'engany segons Wardle com una etiqueta a part, sense que mogui l'índex.
7. **L'efecte en les persones.** Com canvia el que creiem i compartim quan una informació arriba amb el seu índex i les seves proves. Hi ha estudis sobre com avaluem la credibilitat, però encara cap treball propi.

Com citar-lo

Trabazos Clavería, Ó. (2026). *Mesurar la corroboració, no la veracitat: un índex de confiança amb proves per a la premsa digital* (working paper, versió 1.0, sense revisió per parells). https://oscartrabazos.com/papers/trabazos_2026_corroboracion_ca.pdf

Referències

- Ahmed, H. (2017). *Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity*. Tesi de màster, Universitat de Victoria.
- Bowman, S. R., Angeli, G., Potts, C. i 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. i 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. i 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. i de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. arXiv:1704.07506.
- Thorne, J., Vlachos, A., Christodoulopoulos, C. i Mittal, A. (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. *Proceedings of NAACL-HLT 2018*.
- Trabazos, O., Suárez, S., Bori, R. i 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. i Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Consell d'Europa. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

- Wardle, C. (16 de febrer de 2017). Fake news. It's complicated. First Draft. Recull les «quatre P» que va proposar Eliot Higgins i les amplia. https://firstdraftnews.org/articles/fake-news-complicated/

Lectures d'introducció

- Viquipèdia: Eliot Higgins (en castellà), fundador de Bellingcat.
- Viquipèdia: Fake news (en castellà), que resumeix les set categories de Wardle.

Nota sobre les fonts pròpies

La trajectòria (secció 3) es basa en el blog de l'autor (2009-2026), en l'article de la SEPLN del 2014, en documents de la societat NAD (2018-2020), en dos esborranys inèdits de maig del 2018 i en material no publicat del 2023 (l'abstract de la tècnica de l'impostor i l'esquema del prototip). El mètode i l'avaluació (seccions 4 a 6) surten del registre de mesures de GeriAI, on cada cas i cada canvi estan anotats amb data, versió i resultat.