Measuring corroboration, not truth: a trust index with evidence for the digital press

Óscar Trabazos Clavería · Working paper, not peer reviewed · Version 2.0, November 2026 (v2, 2026-11: figures withheld; v1.0, 2 November 2026)

Note on version 2 (November 2026): figures withheld. The figures marked as withheld (the algorithm's thresholds, weights and caps) have been removed from this version to prevent stories from being crafted to score high. The rest of the text is that of version 1.0. Researchers who need them can request them at oscartrabazos.com/contacto.

Abstract

This paper proposes measuring how much corroboration a piece of information has in the digital press, instead of declaring it true or false. It describes GeriAI, a system that takes a headline or a link, finds pieces that report the same thing, groups copies from the same source, decides whether each piece supports, contradicts or says nothing about it, and combines nine signals into a 0-100 index that always comes with its evidence.

The work gathers a line of research that began in 2007 by measuring tone and reputation in what gets published, asked in 2017 «how can we detect a fake news story?», and in 2026 was reframed as «how much trust does it deserve, and on what evidence?». It presents the method, the translation of an unmeasured 2023 proposal (the «impostor technique») into a measured system, and a preliminary evaluation: about ten real cases and a classification test on 251 headline pairs.

The evaluation is preliminary and says so: the weights are provisional and uncalibrated, the cases are examples rather than a sample, and part of the test labels were set by an AI and have not yet been reviewed by a person. The paper separates what has been built and measured from what was proposed and from what remains open.

Keywords: disinformation, corroboration, independent sources, textual inference, trust index, digital press.

About the author

Óscar Trabazos Clavería has been building systems since 2007 that read what is published in the press and online and measure more than words: first the tone and reputation of brands, then irony and emotion, and since 2017 disinformation.

- **Anpro21 and Brand Rain.** At Anpro21 he launched a measure of brand tone in the press and online in 2009, and in 2010 Brand Rain, a reputation measurement tool. Its technical basis was published in the SEPLN journal *Procesamiento del Lenguaje Natural* (2014).
- **TrawlingWeb.** He is its founder and CEO. TrawlingWeb analyses digital press and social networks from dozens of countries, and GeriAI is its AI analysis layer. The system in this paper works on that corpus.

- **Research with universities.** He is co-author, with Pilar Sánchez-García (Universidad de Valladolid) and João Canavilhas (Universidade da Beira Interior), of a study on how the press in Spain and Portugal covers AI, currently under review. In September 2026 he gave the opening keynote of the 1st IBERPERIA International Congress (Covilhã).
- **Writing.** He has published a blog since 2006; the series on disinformation runs from 2017 to 2026.

This working paper is the first time the disinformation line is presented with a method and measurements. What came before were outreach texts, proposals and prototypes, and they are treated as such here.

1. Introduction

The question of this work is how much trust a published piece of information deserves, measured by how much independent sources in the digital press corroborate it over a given period. It does not ask whether it is true.

Deception is not new. What is new is speed: anyone can publish, nobody has to check first, and whatever fits what the audience wants to hear multiplies on its own. Journalism's classic answer still holds: check against independent sources before accepting something. What changes is the scale at which this has to be done.

A binary verdict fails both ways. A true but recent or local fact may have little coverage; a widely copied hoax may have a lot. That is why the system never says «true» or «false»: it gives a 0-100 index, the label of its band (from «insufficient corroboration» to «high corroboration») and every piece that supports or contradicts it, so that the reader can judge.

This paper contributes three things:

1. An explicit method for measuring corroboration: how pieces are found, how independent sources are counted, how support or contradiction is decided and how the signals are combined.
2. The path from a 2023 methodological proposal, which had no figures, to a system in production with every change measured on real cases.
3. A preliminary evaluation with its limits in plain sight.

It does not offer a systematic evaluation: there is no large test set with human-reviewed labels and no comparison with professional fact-checkers. Section 7 details what can and cannot be concluded.

2. Background

The work rests on four ideas from the literature and one rule of the trade. It is not a systematic review: it cites what guided the design.

- **Types of information disorder.** Wardle and Derakhshan (2017) distinguish misinformation, disinformation and malinformation, and seven forms: satire, false connection, misleading content, false context, impostor content, manipulated content and fabricated content. This

framework describes the *type* of deception; the index in this paper does not classify it, it measures something else (corroboration).

- **The sender's motive.** The «four Ps» that Eliot Higgins attributes to disinformation (passion, politics, propaganda and pay), as reported by Wardle (2017), explain why it is fabricated. They are not measured here; they remain context.
- **Text alone is not enough.** Shu et al. (2017) review automatic detection on social media and conclude that content must be complemented by social context: sources, network and spread. Tacchini et al. (2017) classify posts by who interacts with them, not by what they say. Ahmed (2017) shows the reach and the limit of lexical features (n-grams). This work takes the common conclusion: what gives a piece of information away lies mostly outside its text.
- **Claim verification with evidence.** Automatic claim verification (for example FEVER, Thorne et al., 2018) retrieves texts and decides whether they support or refute a claim, with textual inference (Bowman et al., 2015) at its core. GeriAI uses that scheme, but on press from the same period and to measure corroboration, not to issue a verdict.
- **Journalism's oldest rule.** Check against independent sources before publishing. The index turns it into numbers: how many distinct sources, from how many countries, how many support and how many contradict.

The difference from most of the detection literature is the object of measurement. No «false/true» classifier is trained; the available corroboration is counted, qualified and shown.

3. The research line (2007-2026)

The question changed three times: from understanding tone and reputation (2007-2014) to detecting falsehood (2017-2023) and then to measuring trust (2025-2026). On disinformation, everything until 2026 was outreach writing, proposals and prototypes; the first measurements are the ones in this paper.

Year	Milestone	What it brings to this work	Source type
2007-2009	At Anpro21, a first attempt at an algorithm that understood the sentiment of published text; in 2009, a measure of each brand's «tone» in the press and online	The underlying question: measuring what is published, at scale. And the first wall, sarcasm, which machines of the time could not perceive	Blog (2009)
2010-2011	Brand Rain: brand reputation and perception; a study of the 20N Spanish general election with about 15,000 messages in three weeks	Measuring press, blogs and social networks together, and applying it to politics	Blog (2011)
2012	Threat detection project, according to the author commissioned by the Spanish Ministry of the Interior through a partner company that developed defence software	According to the author, the origin of the idea of «detecting the impostor»	Author's account, no document

Year	Milestone	What it brings to this work	Source type
2013	Irony, sarcasm and tone in the measurement of intangibles	Irony as the most visible error in any automatic opinion measurement	Blog (2013)
2014	Article in *Procesamiento del Lenguaje Natural* no. 53 (SEPLN) on NLP and semantic technology at Brand Rain and Anpro21	The technical basis: active listening, entities, relations and complex networks over the press. Not about disinformation	Peer-reviewed publication
2014	Video-reputation project with Ibermática, Verbio, CSA and the BISITE group (Universidad de Salamanca), supported by the Plan Avanza	When text and image do not match, look twice: the idea returns in 2024 with synthetic video	Blog (2014)
2017	«Fake news in politics: a challenge for democracy» (blog)	The disinformation line begins: the problem is spread, not deception itself	Outreach
2017-2020	Jimcrick, an anti-bullying app by the company NAD	Detecting threats in minors' conversations; according to the author, with the same impostor idea	NAD documents (2018-2020)
2018	Series «Politicians, fake news and big data» (2 parts, unpublished until now; signed at the time with a pseudonym)	Presents the algorithm of Sirajudeen et al. (2017): check the source and search whether the title and topic appear elsewhere. The first appearance of external corroboration in the line, as someone else's method	Own draft on others' work
2023	Series «Detecting fake news» I-III: semantic headlines, two-stage classification, Wardle and Higgins	The theoretical framework and the impostor technique (section 3.1)	Outreach + unpublished abstract
2023	Conversational prototype on the TrawlingWeb corpus, tested with Sam Altman's dismissal	Searched the headline in 12 languages and scored with stars. No engine code and no measurements	Prototype + demo videos
2024	«Sora: changing the rules of the disinformation game»	Coherence between text, image and emotion as a signal against synthetic video	Outreach
2025	«Topic safety» for writers	Three axes: reach, credibility («who else is talking about it?») and the sender's authority. The second axis is the index in this paper	Outreach
2026	Series on disinformation as a risk and verification with generative models	The turn: «we are not looking for the truth»; propagation patterns (speed versus volume, homogeneity, source history)	Outreach
2026	GeriAI: corroboration index in production, with a measurement log	The system and evaluation of sections 4 to 6	System + measurement log

Two caveats about this record. First, some 2023 outreach texts promised more than existed (validating «in three minutes» whether something was false, adversarial networks,

blockchain); none of that was built and this paper does not report it as a result. Second, the 2023 impostor-technique abstract places its origin in a 2012 threat detection project (according to the author, commissioned by the Ministry of the Interior through a partner company that developed defence software; the abstract wrongly attributed it to NATO) and in its later use in Jimcrick, an anti-bullying app. That Jimcrick existed and was an anti-bullying app is documented from 2018 to 2020; that it used this technique, and the 2012 project, are undocumented. They therefore appear in the table as the author's account and support no conclusion of this work.

3.1 The impostor technique (2023)

The impostor technique proposes judging a piece of information by who issues it and where it began, not only by what it says. It was set out in an unpublished abstract of November 2023 and popularised in the third part of the «Detecting fake news» series.

The starting idea is to question everyone taking part in a conversation (users, journalists, media, authorities) to find out who is not what they claim to be: the impostor. According to the author, it was born in the 2012 threat detection project and was reused against school bullying. The abstract applies it to disinformation with three components:

1. **Zone 0.** Trace back to the moment, place and source that first covered the topic. Who reported it first and how it spread from there say more than the text of any copy.
2. **Sender authority.** Review the track record of the outlet and of the signing author: whether they already covered that topic, how consistently and with what reputation in that field.
3. **Discriminative and adversarial AI.** One AI classifies and another questions what the first concludes, to reduce false positives.

The theoretical framework combines Wardle, to know *what type* of deception is present, with Higgins, to ask *why* it spreads (passion, politics, propaganda or pay). Sender authority is the practical answer to one of Wardle's forms, impostor content, which gives the technique its name.

Its limit is that it was an unmeasured proposal. The abstract claims the technique «has proven to be an effective tool», but gives no figures; its results section lists future lines and its references section was never written. The text itself acknowledges it had not been compared with other techniques. This work takes the three ideas and puts them to measurement (section 5).

3.2 Which of Wardle's types corroboration can see

Corroboration sees fabricated content well and very little of the rest. Measuring how far a fact is confirmed does not, on its own, say how a piece deceives.

Type (Wardle and Derakhshan, 2017)	What the index can see	Why
Fabricated content	Yes	With no independent sources reporting it, the index cannot exceed WITHHELD (invented control and single source, section 6)
Misleading content	Partly	Rules for different figures, attribution and hedging avoid counting as support what is not; they do not label the deception
False connection	Partly	In each match the title and body are evaluated separately; whether the analysed headline summarises its own body is not checked
Impostor content	Little	A spoofed outlet publishing alone ends up as «a single source»; really detecting it requires sender authority, which is not built
False context	No	If the fact is true, it is corroborated even when presented in a false context
Manipulated content	No	The system only handles text; it does not analyse images or video
Satire or parody	No	It may come out uncorroborated, but the system does not know it is satire

The marked figures are withheld to prevent stories from being crafted to score high. Researchers who need them can request them at oscartrabazos.com/contacto.

The index therefore does not replace Wardle's classification; it complements it. The type of deception could be added as a separate label, without mixing it into the number (section 8).

4. Method

The index is a weighted sum of nine corroboration signals, capped by four fixed rules and accompanied by every piece that supports it. Every threshold and rule was set by measuring real cases; the weights were not.

4.1 Questions

1. How many *independent* sources report the same as the analysed piece in the period, and from how many countries?
2. Of those that report it, how many support it, how many contradict it and how many only cover the topic?
3. How is it spread over time: with a precursor and staggered growth, or all at once?
4. Which part of those answers depends on how the search is done rather than on what has been published?

4.2 Corpus and sources

The corpus is the digital press analysed by TrawlingWeb, queried through its API with keyword syntax. TrawlingWeb operates under the text and data mining (TDM) exception of Article 4 of

Directive (EU) 2019/790: it analyses public sources and cites each piece with its link, but does not reproduce or redistribute the original editorial content. Each piece carries title, text, domain, author, language, outlet country, publication date and the domain's estimated audience. The analysis uses a 7-day period by default (14 at most) around the piece's date, or backwards from today if there is only a headline. Each analysis makes at most WITHHELD queries, one per second.

The input is a headline with a lead or a link. The analysed piece never counts as a source for itself.

4.3 Procedure

1. **Keys.** A local named-entity model obtains people, organisations and places, prioritising those in the headline; rules add figures and dates.
2. **Search.** One context query (the most specific entity with the other keys) and three strict two-word queries. If a query is truncated, the period is split into up to WITHHELD windows so as not to bias towards the oldest pieces.
3. **Matches.** A local multilingual semantic vector model compares each candidate with the headline and lead. Candidates with similarity ≥ WITHHELD that also contain the anchor entity are accepted.
4. **Independence.** Matches are grouped into sources: same root domain, same agency byline, same author or near-identical text (≥ WITHHELD). A hundred copies of a wire story count as one source.
5. **Support or contradiction.** On a sample of WITHHELD pieces, spread by day and by source and with explicit denials first, a local textual inference model decides whether each piece supports, contradicts or is neutral. It evaluates the title and the first three sentences separately, so that a long body does not dilute a negation.
6. **Rules and review.** Deterministic rules cover what the model misses: an explicit denial («denies that», «refutes») contradicts; hedging («could», «allegedly») and attribution («according to», «claims») do not support. Doubtful pieces, those decided by a patch rule and contradictions (except explicit denials) are reviewed by a language model that only classifies and leaves its reason; it does not write text or provide sources.
7. **Scoring.** The nine signals are computed, weighted and the caps applied.

4.4 Signals and weights

Signal	Weight	What it measures
Independent sources	WITHHELD	Number of sources, saturating at WITHHELD
Support versus contradiction	WITHHELD	(support − contradict) / evaluated, rescaled to 0-1
Country diversity	WITHHELD	Distinct countries among the sources, saturating at WITHHELD
Keys in the headline	WITHHELD	Share of keys appearing in each match's title
Audience	WITHHELD	Unique visitors of the domain on a log scale

Signal	Weight	What it measures
Source quality	WITHHELD	Outlet tier; today all «unclassified» (WITHHELD)
Time distribution	WITHHELD	Coverage in at least WITHHELD slices of the period, with a precursor
Similarity	WITHHELD	Mean similarity of accepted matches
First appearance	WITHHELD	Whether it circulated before the analysed piece

The marked figures are withheld to prevent stories from being crafted to score high. Researchers who need them can request them at oscartrabazos.com/contacto.

The index is:

$$I = 100 \cdot \Sigma_{i=1\ldots9}\, w_i\, s_i,\quad \Sigma_i\, w_i = 1,\quad s_i \in [0,1]$$

The weights are **provisional**: they come from expert judgement and are not calibrated on data. So that a true regional story is not penalised for lacking international echo, the saturation of sources, countries and audience depends on the expected scope (regional, national or international), detected in the headline.

4.5 Caps

Condition	The index cannot exceed
Fewer than WITHHELD independent sources	WITHHELD
No evaluated piece supports it	WITHHELD
More pieces contradict than support	WITHHELD
Fewer than WITHHELD of evaluated pieces support it	WITHHELD

The marked figures are withheld to prevent stories from being crafted to score high. Researchers who need them can request them at oscartrabazos.com/contacto.

The caps are rules set by hand from cases (section 6), not the result of statistical fitting. They separate «widely spread» from «corroborated»: a hot topic with many sources that do not confirm the fact cannot score high.

4.6 Output

Each analysis returns the index, the band label (0-25 «insufficient corroboration», 25-50 «weak», 50-75 «moderate», 75-100 «high»), each signal with its weight, the caps applied, the queries made, each piece with its verdict and reason, and the limits of the analysis itself. The labels are about trust, never about truth.

5. From the 2023 proposal to the system

Of the 2023 ideas, corroboration and «one AI classifies, another reviews» are built and measured; zone 0 is half done and sender authority is still not built. Several parts of the

prototype were replaced because, once measured, they did not hold up.

2023 idea	Status in 2026	What changed and why
Find who else reports the same in a large corpus	Built	Search in the analysed digital press with context and strict queries
Count results (threshold of WITHHELD per language)	Replaced	Independent sources are counted: agency syndication inflates the raw count
Match by percentage of words (WITHHELD)	Replaced	Semantic similarity with a measured threshold (WITHHELD); with lexical matching, of 124 pieces reporting the same story in other words only 2 were accepted
One AI classifies, another questions (adversarial networks)	Built, without adversarial networks	Local textual inference decides the clear cases and a language model reviews only the doubtful ones
Zone 0: where and when it was first reported	Half done	Within the period, yes (first-appearance signal). Backwards, experimental and switched off: it lacks a positive case to validate it
Sender authority: track record of outlet and author	Not built	Pending a media quality list; today audience is the stand-in
Translate the headline into 12 languages	Not built	Queries in other languages use only entities; semantic vectors cross languages without translation
Star rating with a «mostly false» label	Discarded	A truth verdict promises more than is measured
Classify the type of deception (Wardle)	Out of scope	It describes the deception, not the corroboration; it could be a separate label
Text-image-emotion coherence	Out of scope	The system only handles text

The marked figures are withheld to prevent stories from being crafted to score high. Researchers who need them can request them at oscartrabazos.com/contacto.

The general lesson is that the ideas survived and the naive implementations did not. Counting results, matching words and issuing a verdict seemed reasonable in a proposal; on real cases, each failed in a concrete, measurable way.

6. Preliminary evaluation

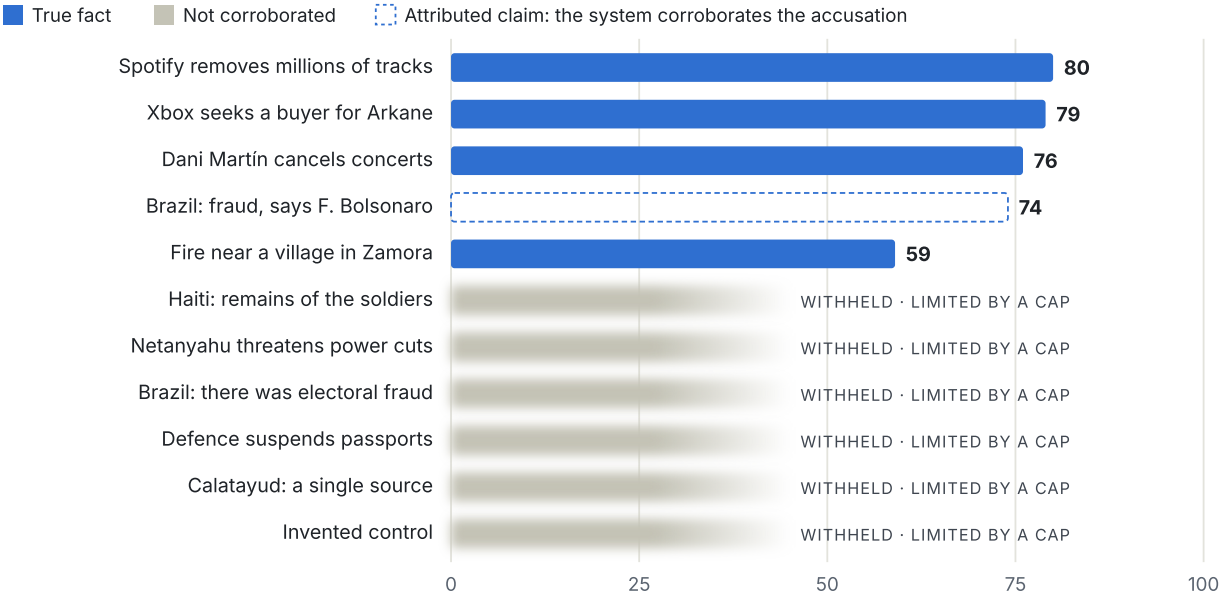
In ten cases measured with the same version, the four true facts scored between 59 and 80 and the six uncorroborated ones stayed at the low end of the scale, with no overlap. This suggests the index orders cases in the expected direction; it is not a measure of accuracy: the cases are few and were chosen by hand.

6.1 Real cases

The cases were chosen to cover different types: widely reported facts, breaking news, regional stories, hoaxes already debunked by fact-checkers, a single-source claim and an invented control. Each was analysed with the period of its publication and at most WITHHELD queries.

True cases score 59-80; uncorroborated ones stay at the low end

Index from 0 to 100 per case, same system version, October 2026



GeriAI measurement log · 11 cases analysed on 30 October 2026 with the same version

The three uncorroborated cases at WITHHELD and the three at WITHHELD are there because of a cap, not because of the sum of signals. Expected scope was added later, and the regional Zamora story, covered by the two local outlets that reported it, rose from 59 to 75; international cases did not change.

Three cases explain how the caps were set:

- **Denial (Haiti).** «The remains found are those of the missing soldiers», which the president of El Salvador denied days later. The first version scored 75: the search only brought the oldest pieces and no denial got in. With windowed search, a sample spread across days and denial rules, it stays at WITHHELD.
- **Single source (Calatayud).** A local story published by a single outlet scored 54 because it «supported itself». Hence the cap at WITHHELD with fewer than WITHHELD sources.
- **Hot topic (Brazil).** «There was fraud in Brazil's elections»: 13 sources and 9 countries, but only 2 of the evaluated pieces supported it. The spread signals added up to 77. Hence the support-share cap: in the true cases that share was WITHHELD to WITHHELD and in the uncorroborated ones 0 to WITHHELD; the WITHHELD threshold falls in the gap.

The documented exception is attribution. «There was fraud in Brazil, *according to a complaint by* Flávio Bolsonaro» scores 74: the system corroborates that he made the complaint, which is

true and widely reported, not that there was fraud. The hypothesis must be the fact, without «according to X».

6.2 Support or contradiction classification

Evaluators were compared on 251 real pairs (the title of a piece against the analysed headline) from the cases above: 156 from a first batch and 95 added later.

Evaluator	Accuracy	False contradictions	Cost of the 251 pairs
Local textual inference, no rules	0.65	—	0
Local textual inference with rules	0.78	7	0
Low-cost language model	0.94	4	0.008 USD

Two results matter more than overall accuracy. First, 20 of the 37 contradictions in the first batch are missed by every inference model: they are denials of the type «X denies that…», which in strict logic do not imply the opposite. A lexical rule solves them, not a larger model. Second, the language model catches all 39 denials without rules.

Caveat on these figures. The labels of the 251 pairs were set by an AI and have not yet been reviewed by a person; the language model evaluated «reasons» much like whoever labelled them, so its 0.94 is probably inflated. Moreover, the test uses titles only, while the system uses title plus three sentences. It serves to compare evaluators with each other, not as expected accuracy.

6.3 The hybrid system on real cases

In six cases with the same sample, review of doubtful pieces by the language model did not move the controls (Spotify, ECB, Dani Martín, Haiti and Brazil stayed the same) and corrected one case: Nvidia went from 75 to 81 because the hedging rule («alleged agent») left as neutral 15 pieces that did confirm the investigation. It reviewed between 2 and 15 pieces per analysis, at a cost of 0.0003 to 0.0016 USD and 0 to 3 extra seconds.

6.4 What real use taught

Once the system was open to the public, real analyses revealed failures the chosen cases had not shown. Three examples: the last day of the period was left out of the search, so a same-day story found none of its echoes (from 0 to 53 once fixed); with link input, entities from the body dragged the search to another topic; and pieces published before the event («pending a vote») were read as contradicting that it happened. Each failure was fixed and measured with one case and two controls.

7. Limits and threats to validity

What this paper allows us to claim is small and specific: the index orders about ten cases in the expected direction, and each rule fixes the failure that motivated it without breaking the controls. It does not allow claiming an accuracy rate or that the index generalises.

- **Small, hand-picked sample.** About ten calibration cases and some twenty real analyses, chosen by type rather than at random. The same cases were used to set rules and to check them, which favours fitting to those cases.
- **Uncalibrated weights.** The nine weights come from expert judgement. Another reasonable allocation could change the order of close cases.
- **Labels without human review.** See the caveat in section 6.2.
- **No external reference.** The index has not been compared with the judgement of professional fact-checkers or with public verification datasets.
- **Corpus dependence.** If a piece is not in the analysed corpus or the search does not find it, the system cannot evaluate it: a 0 may mean «there is nothing» or «it was not found». A fact told in very different words may lose pieces.
- **Noisy dates.** The publication date given by outlets is not always reliable, and the date of the analysed piece is not the date of the event; this limits the time signals.
- **What it does not measure.** It does not detect satire, does not analyse images or video, does not measure the author's authority and cannot tell whether a claim is an opinion or a thesis with no verifiable fact.
- **Corroboration is not truth.** A hoax copied by many independent sources that do not deny it would score high. The index measures how much something is corroborated, and is always presented that way.

8. Open questions

The next necessary step is an evaluation set with human labels; without it, none of the following questions can be settled.

1. **Evaluation with an external reference.** A set of at least 300 pairs with labels reviewed by two people, and a sample of claims already checked by professional fact-checkers to compare the index with their judgement.
2. **Weight calibration.** Fit the nine weights on that set and report how much the order of the cases changes.
3. **Zone 0 backwards.** Find the real first appearance outside the analysed period. It is built and switched off: it lacks a positive case to validate it. One option is to use the date of the first own observation in the continuous archive, more reliable than the date each outlet declares, though it depends on how many months the archive covers.
4. **Sender authority.** Know whether the outlet or author already covered the topic and how consistently, once a media quality list exists.
5. **Date of the event.** Estimate when something happened, not only when the piece was published, to separate what precedes the event from what contradicts it.
6. **Opinions and theses.** Detect headlines with no verifiable fact and warn instead of scoring. Along the same lines, add the type of deception according to Wardle as a separate label that does not move the index.

7. **The effect on people.** How what we believe and share changes when information arrives with its index and its evidence. There are studies on how we judge credibility, but no work of our own yet.

How to cite

Trabazos Clavería, Ó. (2026). *Measuring corroboration, not truth: a trust index with evidence for the digital press* (working paper, version 2.0, not peer reviewed). https://oscartrabazos.com/papers/trabazos_2026_corroboracion_en.pdf

References

- Ahmed, H. (2017). *Detecting Opinion Spam and Fake News Using N-gram Analysis and Semantic Similarity*. Master's thesis, University of Victoria.
- Bowman, S. R., Angeli, G., Potts, C. and 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. and 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. and 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. and de Alfaro, L. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. arXiv:1704.07506.
- Thorne, J., Vlachos, A., Christodoulopoulos, C. and Mittal, A. (2018). FEVER: a large-scale dataset for Fact Extraction and VERification. *Proceedings of NAACL-HLT 2018*.
- Trabazos, O., Suárez, S., Bori, R. and 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. and Derakhshan, H. (2017). *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe. https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c
- Wardle, C. (16 February 2017). Fake news. It's complicated. First Draft. Reports the «four Ps» proposed by Eliot Higgins and extends them. https://firstdraftnews.org/articles/fake-news-complicated/

Introductory reading

- Wikipedia: Eliot Higgins, founder of Bellingcat.
- Wikipedia (Spanish): Fake news, which summarises Wardle's seven categories.

Note on the author's own sources

The research line (section 3) draws on the author's blog (2009-2026), the 2014 SEPLN article, documents of the company NAD (2018-2020), two unpublished drafts from May 2018 and unpublished material from 2023 (the impostor-technique abstract and the prototype outline).

The method and evaluation (sections 4 to 6) come from the GeriAI measurement log, where every case and every change is recorded with date, version and result.