A Multi-Layer AI Framework for Information Landscape Analysis

Maryam Fooladi, Federico Bottino

Kakashi Ventures / Newjee

Abstract

This paper introduces a web-based AI platform for structured analysis of information disorder. Rather than delivering simplified verdicts, the system decomposes content into eleven independently scorable dimensions, including metadata verification, factual claim extraction, source credibility profiling, framing analysis, and propagation detection. At the heart of its architecture sits what we call claim–source independence: factual veracity and source credibility run on separate tracks, preventing the guilt-by-association reasoning that compromises most existing tools. We demonstrate the platform through a case study of the February 2026 Russian disinformation campaign targeting Emmanuel Macron via fabricated Epstein connections. The multi-scale architecture (keyword-level ecosystem mapping paired with URL-level source evaluation) correctly differentiates three qualitatively distinct information problems nested within a single topic space: active fabrication, quality journalism about fabrication, and passive quality degradation through unvetted republication. Each receives its own trust grade, manipulation profile, and source assessment. We are candid about limitations, including the fundamental difficulty of building on LLM-generated outputs without validated evaluation metrics.

Keywords: information disorder, disinformation detection, claim–source independence, multi-layer analysis, LLM-based analysis, FIMI

1. Introduction

False stories on Twitter spread roughly six times faster than accurate ones (Vosoughi et al., 2018). The downstream consequences are not theoretical. Allcott and Gentzkow (2017) traced measurable distortions in electoral behavior, while van der Linden and colleagues (2020) documented how disinformation warped public health responses during COVID-19. But “fake news” as a framing device has always been too blunt. Wardle and Derakhshan (2017) proposed the “information disorder” framework precisely because the actual landscape is far messier than any binary between true and false can accommodate. Their taxonomy forces a reckoning with intent, context, and propagation pathways, all of which matter as much as whether a claim happens to be factually correct.

Traditional fact-checking hits a wall here. Manual verification cannot match the velocity of online content flows. Even if it could, binary verdicts leave enormous analytical territory unexplored. Consider the statement that is technically accurate but strategically omits the one detail that would reverse its meaning. Or the legitimate photograph placed in a fabricated context. These are not edge cases. They are, increasingly, the norm.

This paper presents a web-based AI platform that approaches the problem differently. Instead of delivering verdicts, the system decomposes content into independently scorable dimensions across eleven analytical layers. The central architectural decision is claim–source independence: factual veracity and source credibility are scored on separate tracks, preventing guilt-by-association reasoning where disreputable outlets’ accurate reporting gets dismissed and prestigious broadsheets’ misleading framing gets a pass. The platform operates in two modes: URL-based deep analysis of individual articles and keyword-based narrative mapping across sources. Both produce transparent trust scores aimed at journalists, researchers, and policy analysts who need to understand information ecosystems rather than receive simplified judgments about them.

One clarification is important. The platform is not a primary disinformation detector. It does not crawl bot networks or perform technical attribution. It operates downstream, synthesizing publicly available reporting about

information operations into structured, multi-layered analytical outputs. This distinction shapes how every subsequent result should be interpreted.

2. Related Work

Research on computational approaches to disinformation has expanded rapidly, spanning automated fact-checking, credibility assessment, and narrative analysis. We focus on threads most directly relevant to our contribution; comprehensive surveys are available in Zeng et al. (2021) and Guo et al. (2022).

Automated fact-checking systems, from ClaimBuster (Hassan et al., 2017) through FEVER (Thorne et al., 2018), treat verification as classification: claims mapped to labels of true, false, or intermediate. This works for straightforward assertions but lacks vocabulary for subtler pathologies of information disorder. What do you do with the accurate statement that strategically withholds reversing context? Framing-as-classification has a ceiling, and it is lower than we might like.

Source credibility assessment has taken two paths: human curation (NewsGuard, Media Bias/Fact Check) and automation (Baly et al., 2018; Gruppi et al., 2021). The persistent limitation is that source credibility and claim veracity keep getting entangled. A tabloid can publish careful investigative work. A prestigious broadsheet can frame facts misleadingly. Remarkably few systems treat disentangling these dimensions as a first-order design requirement.

Applied platforms including Bellingcat, VeraAI (Lupi et al., 2023), and AskVera (Wührl et al., 2023) have operationalized research for journalist workflows. These represent real progress. Yet most default to verdict delivery over analytical transparency: telling users what to think about content rather than showing them what is happening in its information ecosystem.

Our platform addresses these gaps simultaneously. Unlike claim-verification systems, it analyzes multiple dimensions without forcing content through a classification bottleneck. Unlike source-rating tools, it explicitly separates claim veracity from source credibility as a concrete scoring mechanism. Unlike tools built exclusively for professional fact-checkers, it makes structured analysis accessible to broader audiences, including researchers investigating cross-cultural information dynamics (Dabbous et al., 2022) and the challenges of annotator disagreement in disinformation labeling (Plank, 2022).

3. System Architecture

3.1 Overview and Transparency

The platform runs on Next.js 14 with a Supabase PostgreSQL backend. The analytical core uses Claude Sonnet 4 (Anthropic), augmented with real-time web search; 20 queries for main analysis, 8 for propagation tracking. Topic classification runs on GPT-4o-mini (OpenAI). These three pipelines execute concurrently.

We should be transparent: this is an LLM-based platform, not a traditional NLP pipeline with trained classifiers and evaluation metrics on held-out test sets. The analytical outputs come from structured prompting. This has implications for interpretability and validation that we address in Section 5.

3.2 Analysis Modes

URL-based analysis performs deep examination of individual articles, extracting claims, evaluating source health, analysing framing and metadata. Keyword-based analysis aggregates across multiple sources to map the narrative landscape around a topic. Their relationship is deliberate: keyword analysis provides the ecosystem-level picture, URL analysis provides source-level granularity. Running both on the same topic produces different outputs, and they are meant to be different.

3.3 Analytical Framework

The framework decomposes content into eleven independently scored layers: (1) Metadata Assessment, examining publication date, authorship, technical authenticity signals; (2) Factual Claims Analysis, with multi-source triangulation; (3) Source Health Scan, evaluating publisher credibility along six weighted axes (Reputation 20%, Factual Consistency 30%, Transparency 15%, Bias History 5%, Manipulation History 20%, Fact-Check Collaboration 10%); (4) Narrative Analysis, tracking information organization and omissions; (5) Media Integrity Assessment, targeting image and video manipulation; (6) Language and Ideological Analysis, drawing on computational pragmatics to surface how language choices shape interpretation

independently of factual content; (7) Framing Detection, maintaining strict separation between descriptive Narrative Frames (characterizing which strategies are employed) and quantitative Manipulation Indicators (assessing how strongly manipulation patterns manifest); (8) Propagation Analysis, tracking spread patterns with outlet-frame mapping; (9) User Interaction Prompts; (10) Reliability Summary; (11) Suggested Actions.

The weighting rationale follows established practice. We weight Factual Consistency highest at 30% because historical accuracy is the strongest predictor of future reliability. Bias History receives only 5%: editorial perspective does not inherently compromise factual accuracy. A source can be openly partisan and scrupulously accurate at the same time.

3.4 Claim–Source Independence

This is the platform's most consequential design decision. Conventional fact-checking conflates source reputation with content accuracy, and both directions produce errors. The system flags high-truth, low-source (HTLS) scenarios, triggered when claim veracity ≥ 0.80 and source reliability ≤ 0.40. Rather than averaging these into a comfortable middle score, the system surfaces the tension explicitly. The inverse case receives analogous treatment.

3.5 Scoring and Output

Trust scores map onto four tiers: A (75–100%), B (50–74%), C (25–49%), D (0–24%). Calibration caps prevent overconfidence, even strong outlets cannot exceed approximately 0.88. Results appear across tabbed views (Facts, Source, Citations, Verdict, Framing, Propagation, Actions, Entities), each with transparent explanations.

4. Case Study: The Macron-Epstein Disinformation Campaign

4.1 Case Selection

In early February 2026, Russian-linked networks launched a Foreign Information Manipulation and Interference (FIMI) operation falsely linking President Macron to convicted sex offender Jeffrey Epstein. The campaign piggybacked on the legitimate release of Epstein court documents by the U.S. Department of Justice (January 30, 2026), mixing real document references with fabricated emails and AI-generated video. France's Viginum agency attributed the operation to networks Storm-1516 and Matryoshka; Bot Blocker and outlets including Euronews, France24, and Reuters independently confirmed attribution.

We selected this case because it represents documented, attributed FIMI with ground truth, deploys multiple disorder types simultaneously, and has European focus and recency aligning with this community's priorities. Critically, the analysis used the keyword query "Macron Epstein" without terms like "disinformation" or "Russia," testing whether the system could surface information disorder unprompted. The analysis ran on February 13, 2026.

4.2 Keyword Ecosystem Analysis

The system assigned Trust Grade D (18%), with Claim Veracity at 5% and Source Reliability at 85%. The divergence matters. The reporting sources (Euronews, France24, Reuters) are highly credible, but the claims circulating in the ecosystem (fabricated emails, doctored documents, AI-generated video) are almost entirely false. The trust formula correctly produces a low composite grade reflecting ecosystem-level falsity rather than inflating the score based on debunking source quality.

The Facts layer extracted seven verified claims, including Viginum's detection and attribution, distribution through a cloned France-Soir website (france-soir.net impersonating france-soir.fr), and identity theft of Le Parisien journalist Victor Cousin.

The Framing layer identified three Narrative Frames (*Scapegoating*, *False Victimization*, *Moral Panic*) alongside five Manipulation Indicators at high confidence: fabricated_news (95%), disinformation_campaign (90%), imposter_source (85%), manipulated_content (80%), influence_operation (90%).

The Propagation layer proved particularly revealing. The Outlet-Frame Mapping (Figure 1) identified three clusters: (1) an *Official Counter-Disinformation Frame* (Euronews, France24, Reuters), (2) a *Technical Analysis Frame* (UNITED24 Media, The Insider, Bot Blocker), and (3) a *Sensationalist Amplification Frame* (Pravda EN, fake France-Soir clones). The ecosystem is tripartite, not binary. Any classifier forcing a two-way split would miss this.

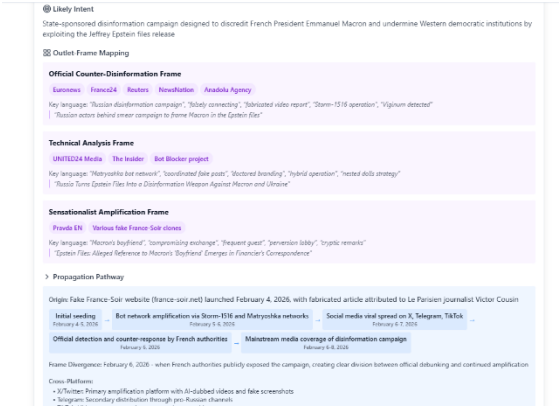

Figure 1: Outlet-Frame Mapping showing three frame clusters with associated outlets and characteristic language.

The **Entity Resolution** of Victor Cousin went beyond standard NER, identifying Cousin as a *victim* of identity theft by cross-referencing factual claims, source analysis, and propagation patterns simultaneously.

4.3 Multi-Scale Comparison

To test multi-scale differentiation, we ran URL-level analyses of two sources within the same ecosystem.

Euronews received Trust Grade A (86%), Claim Veracity 92%, Source Reliability 85%. This is fundamentally different from the keyword analysis's D (18%), and it should be. The keyword analysis evaluates the *ecosystem of claims* (dominated by fabrication); the URL analysis evaluates *one article's reporting* (well-sourced journalism). Same topic, different scale, different valid outputs. The Manipulation Indicators confirmed the distinction: reliable_reporting (85%), government_source_reliance (75%), anti_disinformation_narrative (70%).

IndoPremier covered the legitimate Fabrice Aidan diplomat case (a real French diplomat named in authentic Epstein files). Trust Grade: B (72%), Claim Veracity 82%, Source Reliability 35%. The system triggered an **HTLS conflict flag**: "reliable information from a questionable source." IndoPremier's claims scored high because they originate from AFP wire reporting; the source scored low because it lacks editorial infrastructure for independent verification. A qualitatively different problem from both ecosystem-level fabrication and quality original reporting. Manipulation Indicators confirmed this third profile: republished_content (85%), limited_verification (70%), neutral_framing (60%).

4.4 Synthesis

Figure 2 summarizes the three-scale comparison.

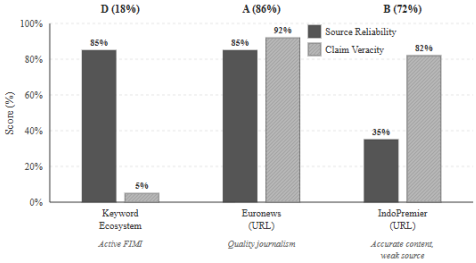

Figure 2: Comparative analysis of Source Reliability and Claim Veracity across three analytical scales within the "Macron Epstein" information ecosystem. Trust grades shown above each group. The Keyword Ecosystem analysis reveals a sharp divergence between high Source Reliability (85%) and near-zero Claim Veracity (5%), reflecting ecosystem-level falsity despite credible reporting sources. The IndoPremier analysis triggers a high-truth, low-source (HTLS) conflict flag, with Claim Veracity (82%) substantially exceeding Source Reliability (35%).

The same topic contains three qualitatively distinct information problems, and the architecture correctly differentiates all three. Keyword analysis alone would not reveal that sources range from excellent to questionable. URL analyses alone would miss the ecosystem-level FIMI operation. The scales are complementary, not redundant.

The Trust Score paradox D grade despite 85% Source Reliability, illustrates what happens when source quality and claim quality are honestly integrated. The Framing dual structure allows describing how actors frame issues independently from assessing whether manipulation is present. The Outlet-Frame Mapping reveals a tripartite ecosystem invisible to binary bias classifiers. The Entity Resolution of Victor Cousin demonstrates what becomes possible when multiple layers converge on the same entity.

5. Discussion and Limitations

Analytical positioning. The platform did not discover the Russian bot networks. Viginum and Bot Blocker did that work. What the platform did was synthesize fragmented public reporting into structured outputs (frame

clusters, propagation pathways, severity-graded patterns, cross-referenced entity profiles) that would otherwise require considerable manual effort. The value proposition is structuring intelligence, not generating it.

The multi-scale finding is the paper's central contribution. What appears to be a single topic actually contains three qualitatively distinct information problems: active fabrication, responsible reporting about fabrication, and passive quality degradation. This validates offering both ecosystem-level and source-level analysis. The case study is itself inherently cross-cultural: a Russian operation targeting a French president, reported by European and Turkish outlets, with an Indonesian website republishing French wire content. The platform's English-language retrieval captured only one slice of this multilingual landscape, underscoring the urgency of multilingual expansion for any system analyzing European FIMI operations where campaigns exploit language boundaries (Dabbous et al., 2022).

Limitations. Every quantitative output (claim veracity, source reliability, manipulation indicators, trust grades) comes from Claude Sonnet 4 through structured prompting. None have been validated against ground truth annotations or benchmarked against expert judgments at scale. The scores produce plausible orderings, but calibration properties are unknown. A 92% veracity score does not carry the same meaning as a 92% F1 from a validated classifier. The source health weights are heuristic. The system retrieves only English-language sources. Building on an LLM means inheriting its failure modes.

A preliminary evaluation by two computational linguistics researchers showed 67% agreement across nine assessment dimensions, with full agreement on the keyword ecosystem and disagreements on cases where claim–source independence creates analytical tensions. These disagreements illustrate the human-label variation problem (Plank, 2022) applied to information disorder assessment: reasonable experts applying different weighting intuitions to identical outputs.

6. Conclusion

We have presented a multi-layer AI platform for structured analysis of information disorder at two complementary scales. The case study demonstrated that a single topic space can harbor qualitatively distinct information problems, each correctly characterized with distinct trust grades, manipulation profiles, and source assessments. The platform occupies a specific niche: structuring fragmented reporting into multi-layered outputs that support human analysis. Its limitations (unvalidated confidence scores, English-language bias, heuristic weights, no user studies) define the research agenda: multilingual retrieval, longitudinal tracking, empirical validation, and structured user studies with information professionals.

7. Ethical Considerations

This work raises several ethical questions that we want to address directly.

The platform relies on commercial LLMs (Claude Sonnet 4 and GPT-4o-mini) for its analytical core. This creates a dependency on opaque, proprietary systems whose internal reasoning cannot be fully audited. Users of the platform must understand that trust scores, manipulation indicators, and veracity assessments are probabilistic outputs generated through structured prompting, not deterministic classifications from validated models. We have attempted to mitigate potential overreliance by making the scoring transparent and by presenting analytical decompositions rather than binary verdicts.

The case study involves a real, documented disinformation campaign. All information referenced in this paper comes from publicly available reporting by established outlets (Euronews, France24, Reuters) and official government sources (Viginum). We do not reproduce or amplify any of the fabricated content from the campaign itself. The identity theft of journalist Victor Cousin is discussed because it was already publicly reported and attributed; we include it to demonstrate the platform's entity resolution capabilities, not to further expose the victim.

There is a broader risk that platforms performing automated information analysis could themselves become tools of censorship or political manipulation if deployed without safeguards. Our system is designed to support human analysts, not to replace editorial judgment or automate content moderation decisions. We strongly discourage any use of this platform for automated content removal or suppression without human oversight.

Finally, the platform currently operates only in English, which introduces a linguistic bias in its analytical outputs. Information ecosystems are inherently multilingual, and evaluating them through a single language inevitably produces an incomplete picture. We acknowledge this limitation and consider multilingual expansion a priority for responsible development.

Bibliographical References

Allcott, H. and Gentzkew, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2):211–236.

Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. (2018). Predicting Factuality of Reporting and Bias of News Media Using Twitter User Stances. In *Proceedings of ACL*, pages 163–168.

Dabbous, A., Aoun Barakat, K., and de Quero Navarro, B. (2022). Fake News in the Middle East: A Cross-Cultural Comparison. *International Communication Gazette*, 84(7-8):640–660.

Gruppi, M., Horne, B. D., and Adalı, S. (2021). NeLa-GT-2020: A Large Multi-Labelled News Dataset for The Study of Fake News. In *Proceedings of ICWSM*, 15(1):892–897.

Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A Survey on Automated Fact-Checking. *TACL*, 10:178–206.

Hassan, N., Li, C., and Tremayne, M. (2017). ClaimBuster: The First-Ever End-to-End Fact-Checking System. *Proceedings of the VLDB Endowment*, 10(12):1945–1948.

Hussain, M. N., Tokdemir, S., Agarwal, N., and Al-Khateeb, S. (2025). Fake News Datasets: A Comprehensive Survey. *Information Processing & Management*, 62(1):103551.

Lupi, M., Cavalcanti, G., and Caselli, T. (2023). VeraAI: A NLP System for the Detection of Online Disinformation. In *Proceedings of PoliticalNLP at LREC-COLING 2024*, pages 1–10.

Plank, B. (2022). The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proceedings of EMNLP*, pages 10671–10682.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: A Large-Scale Dataset for Fact Extraction and VERification. In *Proceedings of NAACL-HLT*, pages 809–819.

van der Linden, S., Roozenbeek, J., and Compton, J. (2020). Inoculating Against Fake News About COVID-19. *Frontiers in Psychology*, 11:566790.

Vosoughi, S., Roy, D., and Aral, S. (2018). The Spread of True and False News Online. *Science*, 359(6380):1146–1151.

Wardle, C. and Derakhshan, H. (2017). Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. *Council of Europe Report* DGI(2017)09.

Wührl, A., Sander, M., Müller, J., and Klinger, R. (2023). AskVera: Informed Consent for News Credibility Assessment. In *Proceedings of EMNLP: System Demonstrations*, pages 386–396.

Zeng, X., Abujabal, A., Saha Roy, R., and Weikum, G. (2021). Automated Fact-Checking: A Survey. *Knowledge and Information Systems*, 63(7):1675–1714