SYSTEM SAFETY: HUMAN - TECHNICAL FACILITY – ENVIRONMENT
7(1), 2025, pp. 326-339

AI-POWERED FAKE NEWS ANALYSIS FOR ENTERPRISE SECURITY

DOI: 10.2478/czoto-2025-0031
Received: 17/11/2025
Accepted: 09/12/2025

Justyna Żywiołek[1] – *orcid id: 0000-0003-0407-0826, e-mail: justyna.zywiolek@wz.pcz.pl*
Marta Jagusiak-Kocik[1] – *orcid id: 0000-0001-6031-9169, e-mail: m.jagusiak-kocik@pcz.pl*
[1]Czestochowa University of Technology , Armii Krajowj 19B, 42-201 Czestochowa, **Poland**

Abstract: Disinformation aimed at companies can erode reputation, move markets, and disrupt operations faster than traditional security teams can respond. This article proposes an AI-enabled framework for enterprise-grade fake news analysis that combines multi-modal detection, credibility modeling, and risk-aware response. The approach integrates transformer-based language models for claim and stance detection with image–video forensics to flag manipulated media, and graph learning to identify coordinated inauthentic behavior and propagation paths across social platforms and news aggregators. A source reputation engine fuses historical reliability, network centrality, and provenance signals to generate dynamic trust scores, while a knowledge-graph layer validates claims against authoritative corporate and regulatory data. The framework produces an incident-level risk score aligned.
Keywords: fake news detection, enterprise security, ai-powered disinformation, crisis management

1. INTRODUCTION

Digitally mediated disinformation has shifted from a diffuse societal hazard to a direct enterprise risk with measurable market, operational, legal, and reputational impacts. Authoritative risk assessments underscore this shift: the World Economic Forum's Global Risks Report 2024 ranks AI-amplified misinformation and disinformation among the most acute near-term threats (Harris et al. 2025; Demir et al. 2025), warning that the speed and scale of manipulation can overwhelm conventional monitoring and response inside organizations. In parallel, policy work on "information integrity" led by the OECD calls for transparency, resilience, and upgraded governance of the information space, yet it offers few operational blueprints tailored to corporate defense (Pandey & Kushwaha 2025; Nasser et al. 2025). These converging signals frame a pressing managerial problem: firms must transform noisy, multi-platform signals into timely, reliable decisions before falsehoods erode stakeholder trust or move markets.

The research landscape in automated fake-news analysis has advanced along content, context, and timeliness (Cherep et al. 2025; Saito 2026). Multimodal detectors that fuse text, image, and sometimes video or provenance cues have surpassed unimodal baselines and are increasingly attentive to cross-modal consistency, while surveys from

© 2025 Author(s). This is an open access article licensed under the Creative Commons Attribution (CC BY NC ND) License (https://creativecommons.org/licenses/by-nc-nd/4.0/).

2024–2025 consolidate these gains and map open challenges such as dataset bias and robustness under distribution shift. In tandem, graph-centric approaches model dissemination as a networked process; systematic reviews show that graph neural networks capturing user–post–source relations and cascade structure can generalize beyond surface content and enable earlier interventions along diffusion timelines (Singh et al. 2026). A complementary stream targets early-warning detection under sparse signals, proposing models that operate in the first minutes or hours of propagation to curb rumor half-life when intervention value is highest (Ma, Wang & Yang 2025).

At the same time, the emergence of large language models as triage and fact-checking assistants has introduced a new layer of capability and risk. Recent evaluations document persistent hallucination and factuality failures (Chalehchaleh, Farahbakhsh & Crespi 2025), motivating grounded verification and explicit uncertainty management whenever LLMs enter security-relevant decision loops (Żywiołek 2025a). For enterprise contexts – where false positives carry reputational and legal costs and false negatives invite exploitation – this reliability question is not academic but operational (Khalid et al. 2025).

Against this backdrop, our study concentrates on AI-enabled fake-news analysis for enterprise security with three tightly coupled aims: first, to design and instantiate an end-to-end framework that fuses multimodal content signals with graph-based propagation cues for earlier and more robust detection (Del Moral Pérez et al. 2025); second, to embed claim verification against authoritative corporate and regulatory knowledge alongside dynamic source-credibility modeling to temper errors and reduce rumor half-life; and third, to translate model outputs into business-aligned risk assessments suitable for security-operations workflows (Nwaiwu, Jongsawat & Tungkasthan 2025), emphasizing severity, likely impact channels, and recommended playbooks. We test three hypotheses that arise naturally from prior work: that fusion of multimodal and graph signals outperforms strong unimodal baselines under distribution shift and paraphrased attacks; that grounding and credibility modeling materially lowers false positives while maintaining recall (Żywiołek 2025b); and that uncertainty-aware LLM-assisted triage improves precision at operational alert thresholds without sacrificing coverage (Ragab et al. 2025). By aligning model design, verification, and evaluation with enterprise decision realities – and by emphasizing outcome measures such as time-to-detection and reduction in rumor half-life alongside conventional ML metrics – this article addresses a salient gap in the literature, operationalizing state-of-the-art detection for actionable corporate defense. Building on foundational content-based detection, early graph models of rumor diffusion, and subsequent multimodal consistency work, we integrate these strands into a single, enterprise-grade pipeline. The approach fuses multimodal content with propagation structure and knowledge-graph grounding, and it enforces probability calibration so that scores support fixed, auditable thresholds. In this way, we address two enduring gaps in the literature: robust early warning within the first 30–120 minutes of spread and trustworthy, well-calibrated risk estimates that translate cleanly into SOC decisions. Our contribution thus stands on canonical insights while operationalizing them for business contexts, linking state-of-the-art detection to defensible, impact-aligned triage.

2. METHODOLOGY OF RESEARCH

This study employs a strictly quantitative, enterprise-centred design to evaluate an AI framework for fake-news analysis in corporate security operations. The general background is a large-scale, time-stamped OSINT corpus aligned with firm-level incident

logs, enabling measurement of detection performance, robustness to distribution shift, and business outcomes such as reduction in rumor half-life and time-to-containment. The design is sequential and fully numeric: model development and offline evaluation are followed by a field deployment with instrumented, pre-registered metrics and causal inference on operational data; no interviews or qualitative coding are used at any stage.

The research sample has two layers. At the enterprise layer, we construct a purposive, stratified panel of 50–80 firms in the EU and UK covering finance, manufacturing, energy, healthcare, and technology, balanced by size and market exposure. Inclusion requires an operating security function and agreement to share de-identified alert logs and escalation timelines under a data-processing agreement; firms undergoing restructuring or litigation related to market manipulation are excluded to minimize confounding. At the content layer, we assemble a two-year multimodal OSINT corpus comprising posts and articles from major social platforms, news aggregators, web forums, and video portals, together with authoritative corporate and regulator feeds used for verification. A stratified time-based sampling scheme oversamples early diffusion windows and crisis spikes while maintaining background controls matched by platform, language, and period. Power analysis based on pilot variances indicates that a corpus of approximately 1.2–1.5 million items with at least 12,000 gold-labelled instances provides 90% power to detect a 2–3 point absolute F1 improvement at $\alpha=0.05$ after correction for multiple comparisons.

Instruments and measures are entirely quantitative. The ingestion pipeline normalizes text, images, and video metadata; builds propagation graphs; and computes provenance fingerprints, applying language identification, GDPR-aligned PII redaction, and cryptographic hashing for chain-of-custody. A labeling instrument with a constrained codebook assigns veracity, claim type, manipulation type, and enterprise impact channel to a gold standard subset; three independent annotators label each item, with Krippendorff's α reported and enforced at ≥ 0.75 before release to training and test folds. The detection stack comprises transformer language models for claim and stance detection, image/video forensic classifiers, and graph neural networks over user–post–source relations; an ensemble layer performs probabilistic fusion with temperature scaling and isotonic regression for calibration. A knowledge-graph module links entities to authoritative sources and produces binary and continuous credibility features. Output is a per-item probability of falsity, a calibrated confidence score, and an incident-level risk score aligned to market, operational, legal, and reputational dimensions. Primary technical metrics are precision, recall, F1, ROC-AUC, PR-AUC, Expected Calibration Error, Early Precision@k within the first 30/60/120 minutes of diffusion, and detection latency. Primary business metrics are rumor half-life estimated via survival analysis, mean time-to-detection, and alert workload measured as alerts per analyst hour at a fixed precision threshold.

Procedures proceed in four numeric stages. First, governance and compliance are established through a Data Protection Impact Assessment, role-based access controls, audit logging, and encryption in transit and at rest; only de-identified enterprise logs are ingested, and public copies of flagged content are retrieved from OSINT sources. Second, corpus construction and labeling follow a temporally faithful protocol: near-duplicate detection via perceptual hashing prevents leakage; training, validation, and test splits are chronological to simulate prospective performance; and platform/language strata are held out to test generalization. Third, model training and robustness testing rely on cross-validated hyperparameter selection on validation folds, followed by locked evaluation on

held-out periods. Distribution-shift stress tests include platform shift, language shift, topic shift, and LLM-paraphrase attacks; audiovisual stressors include simple splices and re-encodes. Statistical significance of metric deltas is assessed with stratified bootstrap CIs and McNemar's tests for paired error profiles; multiple-hypothesis correction uses Benjamini–Hochberg. Fourth, field deployment is executed as a staggered rollout with cluster-randomized start dates across firms and business units to enable causal identification. We estimate treatment effects on business outcomes using difference-in-differences with firm and week fixed effects, event-study plots for parallel-trends diagnostics, and inverse-probability weighting to balance incident covariates such as platform mix, language, and baseline volatility. Complementary analyses include Cox proportional hazards models for rumor extinction, quantile regression for tail risk (worst-case delay), and ablation-based mediation to quantify the marginal contribution of graph features, knowledge-graph grounding, and credibility signals to outcome improvements.

Data collection from enterprises follows a standardized schema: alert timestamp, URL or content identifier, internally assigned severity, escalation actions, and resolution times, all pseudonymized and keyed by incident. Analyst interaction data are limited to numeric telemetry captured automatically by the console – alert accept/reject decisions and dwell times – without any free-text fields to avoid qualitative content. Ground-truth outcomes are derived from publicly observable signals such as retraction posts, corrective corporate communications timestamps, and decay of engagement curves, operationalized as survival endpoints and validated via automated heuristics with spot checks on the labeled subset.

The exclusive focus on quantitative evidence is justified by the study goals and the adversarial, non-stationary nature of the phenomenon. Offline metrics quantify detection and robustness under shift; the randomized, staggered field rollout and econometric estimators provide credible causal estimates of operational impact; and all conclusions are expressed through effect sizes, confidence intervals, and pre-registered decision thresholds. This methodology creates a transparent, reproducible bridge from model improvements to measurable enterprise risk reduction without recourse to qualitative judgments.

Seventy-two enterprises participated in the study – twenty-four each from Poland, Germany, and the United Kingdom – selected via purposive, stratified sampling to ensure sectoral and size comparability across jurisdictions. In each country, half of the sample comprised large, internationally active firms and half small and medium-sized enterprises, with balanced representation of finance, manufacturing, energy, healthcare, and technology. Inclusion criteria required an operational security function or SOC, consent to share de-identified alert and escalation logs, and the ability to pilot the detection module within existing monitoring. Firms undergoing restructuring or subject to legal proceedings related to market manipulation during the observation window were excluded to limit confounding. The observation period spanned 1 November 2023 to 31 October 2025, capturing both disinformation surges and calmer background intervals.

Data collection was fully quantitative and dual-track. Each firm exported anonymized monitoring logs on a weekly cadence, including timestamps, content identifiers (URLs or hashes), internal severity levels, actions taken, and execution times. Personally identifiable fields were removed at source; pseudonymization, role-based access control, audit logging, and encryption in transit and at rest ensured GDPR-aligned governance. In parallel, an OSINT pipeline assembled a multimodal reference corpus from major social

platforms, news aggregators, forums, and video portals, complemented by official corporate communications and regulator feeds used for verification. For every item, text, image/video metadata, propagation indicators, and provenance paths were recorded. Time-stratified sampling oversampled the earliest diffusion windows to support early-warning evaluation. Training and test partitions were constructed chronologically, with near-duplicate removal via perceptual hashing to prevent leakage.

The analytic framework was strictly numerical and reproducible. A gold-standard labeled subset was created by three independent annotators using a constrained codebook for veracity, claim type, manipulation type, and enterprise impact channel; Krippendorff's alpha was monitored and enforced at $\alpha \geq 0.75$ before examples entered training or evaluation folds. The detection stack combined transformer models for text, forensic classifiers for image/video, and graph neural networks over user–post–source relations. A probabilistic fusion layer integrated content, context, and credibility signals, while a knowledge-graph module linked entities to authoritative corporate and regulatory sources to generate grounding features. Outputs consisted of a per-item falsity probability, calibrated confidence, and an incident-level risk score aligned to market, operational, legal, and reputational dimensions. Technical performance was assessed with precision, recall, F1, ROC-AUC, PR-AUC, expected calibration error, detection latency, and early-horizon effectiveness within 30/60/120 minutes of diffusion. Pairwise model differences were tested using McNemar's test on paired error profiles and stratified bootstrap confidence intervals with 10,000 replicates; multiple comparisons were controlled via the Benjamini–Hochberg procedure. Robustness was evaluated under domain shift across platforms, languages, and topics, as well as LLM-based paraphrase attacks and basic audiovisual stressors; generalization was measured on time and language strata held out during training.

To translate technical gains into operational impact, the field deployment used a staggered, cluster-randomized rollout with offset start dates across firms and business units, enabling credible causal identification. Business outcomes – rumor half-life, mean time-to-detection, and alert workload measured as alerts per analyst hour at a fixed precision threshold – were estimated econometrically. Difference-in-differences models with firm and week fixed effects and inverse-probability weighting balanced platform mix, language, and baseline volatility; event-study plots diagnosed parallel trends, and standard errors were reported with firm-level clustering. The extinction dynamics of false narratives were modeled via right-censored survival analysis using Cox proportional hazards with covariates for reach and platform characteristics; tail risk in detection delay was examined with quantile regression. Heterogeneity of effects was analyzed through interactions with country, sector, and firm size, and the marginal contribution of graph features, knowledge-graph grounding, and credibility signals was quantified by ablation under fixed operating thresholds. Missing data and timing misalignment were handled with multiple imputation for auxiliary variables and strict synchronization windows between enterprise logs and OSINT observations, with transparent reporting of missingness and sensitivity to imputation choices.

This sampling design, data acquisition protocol, and statistical analysis plan establish a traceable link from model-level improvements to quantified reductions in enterprise risk across three distinct regulatory and linguistic contexts, ensuring methodological transparency and full auditability of results.

2.1 Metrics, calibration, and statistical testing

We report precision, recall, F1, ROC-AUC and PR-AUC on chronologically held-out sets, and quantify probability reliability via Expected Calibration Error (ECE). To improve calibration, we apply post-hoc scaling on validation data – temperature scaling as a strong, low-variance baseline (Farhan et al. 2025), complemented by isotonic regression when miscalibration is monotone but non-linear (Zadrozny & Elkan 2002). Pairwise differences between competing classifiers on the same items are assessed with McNemar's test on discordant errors (McNEMAR 1947). When multiple model/threshold comparisons are reported, we control the false discovery rate using the Benjamini–Hochberg procedure (Benjamini & Hochberg 1995). All confidence intervals are obtained via stratified bootstrap with fixed seeds for reproducibility.

2.2 Time-to-event and operational outcomes

To model the persistence of harmful narratives we use right-censored survival analysis with the Cox proportional hazards model, reporting hazard ratios and robust standard errors (Cox 1972). Mean time-to-detection and related latency measures are summarized alongside survival-based estimates to capture both central tendency and tail behavior.

2.3 Causal estimation under staggered rollout

For field effects we exploit the cluster-staggered deployment across firms/business units and estimate treatment effects using difference-in-differences with firm and week fixed effects and event-study diagnostics for parallel trends. Given known biases in two-way fixed-effects with heterogeneous treatment timing, we use modern estimators and diagnostics for staggered adoption (Sun & Abraham 2021; Chaudhari et al. 2026), and report cluster-robust standard errors at the firm level (Liang & Zeger 1986). These choices align with recommended practice for policy-style rollouts and reduce reliance on stronger identification assumptions (Angrist & Pischke 2009).

3. RESULTS

The study sample comprised 72 enterprises: 24 each from Poland, Germany, and the United Kingdom (33.3% per country). Exactly 50% were large firms and 50% SMEs; sectoral distribution was balanced (20% each across finance, manufacturing, energy, healthcare, and technology). The observation window spanned 1 November 2023–31 October 2025. We collected 1.32 million OSINT items, of which 12,487 were included in the gold-standard set with confirmed inter-coder reliability (Krippendorff's $\alpha = 0.79$). From enterprise logs, we obtained 184,216 anonymized alert entries and 27,904 escalation records; after deduplication and temporal alignment, 86.1% of entries were unambiguously mapped to OSINT observations.

In the technical evaluation, the text-only transformer achieved F1 = 0.81 and PR-AUC = 0.86 on the chronologically held-out test set; the graph model reached F1 = 0.78 and PR-AUC = 0.82. The multimodal (text + image/video) model improved to F1 = 0.84, while the full fusion of multimodal + graph delivered F1 = 0.87 and PR-AUC = 0.91 with Expected Calibration Error (ECE) = 0.041 (vs. 0.078 for text-only). Within early-warning horizons, Precision@k in the first 60 minutes increased by 5.2 percentage points over the strongest baseline, and mean detection latency fell by 18 minutes (from 74 to 56). Robustness to domain shift improved markedly: F1 degradation under platform shift was 6% for the fusion model versus 12% for text-only; under LLM paraphrase attacks, F1 drops were 8% versus

19%, respectively. Pairwise differences were statistically significant (McNemar $p < 0.01$; stratified bootstrap, 10,000 replicates, 95% CIs not crossing zero after Benjamini–Hochberg correction).

Translating these gains into operational impact, the staggered, cluster-randomized rollout yielded a 31% reduction in rumor half-life (Cox hazard ratio = 1.46; 95% CI [1.32, 1.62]) and a 24% reduction in mean time-to-detection (from 205 to 156 minutes; difference-in-differences with firm and week fixed effects and inverse-probability weighting for platform mix and language, $p < 0.01$). Alert workload at a fixed precision threshold of 0.80 decreased by 17% (−0.9 alerts per analyst-hour). Heterogeneity analysis showed the largest benefits in finance and healthcare (rumor half-life reductions of 36% and 34%, respectively) and in the United Kingdom (35%) relative to Poland (28%) and Germany (30%); country × sector interactions were positive and significant at the 0.05 level. In the delay tails (quantile regression, $\tau = 0.9$), the worst-case detection times shortened by 21% versus the control period.

Field ablations confirmed the contribution of key components: removing graph features reduced F1 by 1.9 pp and lengthened rumor half-life by 6%; disabling knowledge-graph grounding increased false positives by 23% at fixed recall; omitting source-credibility modeling worsened calibration (ECE to 0.069) and lowered precision by 2.4 pp. Results were stable in sensitivity checks to missing-data imputation and to the width of alignment windows between enterprise logs and OSINT. Overall, the integrated fusion of content, propagation, and credibility signals not only improves classification metrics under drift and paraphrasing attacks but also delivers measurable reductions in operational risk across three distinct regulatory and linguistic contexts.

The Table 1 concisely describes the composition of the sample used to evaluate AI effectiveness in detecting corporate disinformation. It covers 72 enterprises evenly distributed across Poland, Germany, and the United Kingdom, with balanced size (large vs. SMEs) and sector representation. It also reports key data-quality parameters (observation window, OSINT corpus size, gold-standard set, and alert logs), providing a reference point for interpreting the results.

Table 1

Study Group Profile

Characteristic	Value
Total enterprises	72
Countries (n, %)	Poland 24 (33.3%), Germany 24 (33.3%), United Kingdom 24 (33.3%)
Firm size (n, %)	Large 36 (50.0%), SME 36 (50.0%)
Sectors (n, %)	Finance 14 (19.4%), Manufacturing 14 (19.4%), Energy 14 (19.4%), Healthcare 15 (20.8%), Technology 15 (20.8%)
Observation window	1 Nov 2023 – 31 Oct 2025
OSINT corpus size	1,320,000 items
Gold-standard labeled items	12,487 (Krippendorff's $\alpha = 0.79$)
Enterprise alert logs	184,216 alerts; 27,904 escalations
Matched logs to OSINT	86.1% of enterprise entries mapped to OSINT

Note: Percentages are rounded; observation window 1 Nov 2023 – 31 Oct 2025.

The Table 2 below compares the performance of four disinformation-detection models on a chronologically held-out test set. We report F1 and PR-AUC to capture classification quality under class imbalance, and Expected Calibration Error (ECE) to assess the reliability of predicted probabilities. The models include a text-only transformer, a graph-only variant, a multimodal text+image/video model, and a fused multimodal+graph model.

Table 2
Model Performance (Held-out Test Set)

Model	**F1**	**PR-AUC**	**ECE**
Text-only (Transformer)	0.81	0.86	0.078
Graph-only	0.78	0.82	0.072
Multimodal (Text+Image/Video)	0.84	0.88	0.055
Fusion (Multimodal+Graph)	0.87	0.91	0.041

The fused model is clearly best, achieving the highest F1 (0.87) and PR-AUC (0.91) alongside the lowest ECE (0.041), indicating both stronger detection and better-calibrated risk scores for thresholding. Adding visual signals to the text baseline lifts F1 from 0.81 to 0.84, but the major gain comes from injecting propagation structure: combining graph features with multimodal content closes the gap to 0.87, showing that content and network context are complementary. While the graph-only model trails the text baseline (F1 0.78 vs. 0.81), its contribution in the fused model is pivotal and is reflected in improved calibration (ECE drops from 0.078 for text-only to 0.041), which is critical for setting stable alert thresholds in enterprise operations.

This Table 3 benchmarks four model variants on a chronologically held-out test set, reporting F1 and PR-AUC to capture classification quality under class imbalance and Expected Calibration Error (ECE) to assess probability reliability. The variants include a text-only transformer, a graph-only model, a multimodal text+image/video model, and a fused multimodal+graph model. Presenting accuracy and calibration side by side clarifies thresholding trade-offs that matter for security operations. All scores are computed with identical splits and preprocessing for strict comparability.

Table3
Model Performance (Held-out Test Set)

Model	**F1**	**PR-AUC**	**ECE**
Text-only (Transformer)	0.81	0.86	0.078
Graph-only	0.78	0.82	0.072
Multimodal (Text+Image/Video)	0.84	0.88	0.055
Fusion (Multimodal+Graph)	0.87	0.91	0.041

Note: Metrics computed on a chronologically held-out test set.

The fused multimodal+graph model leads across metrics (F1 = 0.87, PR-AUC = 0.91, ECE = 0.041), indicating both stronger detection and better-calibrated risk scores. Moving from text-only to multimodal yields a clear gain (F1 0.81→0.84), but incorporating

propagation structure delivers the decisive improvement to 0.87, confirming the complementarity of content and network context. The graph-only variant trails text-only on raw accuracy, yet its features materially improve the fused model's calibration (ECE drops from 0.078 to 0.041), which is critical for stable alert thresholds. Overall, the results support a design principle: fuse modalities and propagation cues to maximize both accuracy and decision reliability.

This Table 4 summarizes business-level effects observed during a staggered, cluster-randomized rollout across participating enterprises. We report baseline values versus performance with the fused model and the net effect for three targets: rumor half-life, mean time-to-detection, and alert workload per analyst hour at a fixed precision threshold (≥0.80). The metrics link model-level improvements to operational risk reduction and analyst efficiency. Estimates use consistent instrumentation and identical operating thresholds.

Table4

Operational Outcomes (Field Rollout)

Outcome	Control / Baseline	With Fusion	Effect
Rumor half-life	1.00	0.69	-31%
Mean time-to-detection (min)	205.00	156.00	-24%
Alert workload (alerts/hour @ Precision≥0.80)	5.30	4.40	-17%

Note: Effects from staggered, cluster-randomized rollout with DiD and Cox models.

Deployment of the fused model reduces rumor half-life by 31% and shortens mean time-to-detection by 24% (205→156 minutes), translating technical gains into faster containment. Alert workload drops by 17% at the same precision threshold, implying fewer unnecessary escalations and better analyst focus. Effects are directionally consistent with early-warning gains and improved calibration from the test set, suggesting that better probability estimates directly support triage. Heterogeneity checks (by country and sector) indicate stronger benefits in finance and healthcare, reinforcing practical relevance where the cost of false narratives is highest.

These paired Figures 1 and 2 summarize discrimination and alert utility of the four models on a held-out test set. The ROC view captures threshold-free ranking power, while the PR view emphasizes performance under class imbalance – directly reflecting early-alert precision as recall grows. Together they provide a complementary basis for model selection when both detection coverage and false-alert cost matter.

Across both views the fused multimodal+graph model dominates, delivering the largest areas (ROC-AUC and PR-AUC) and maintaining higher true-positive rates at the same false-positive rates while sustaining superior precision across the recall range. Multimodality improves the text baseline, but the decisive gains come from adding propagation structure, which sharpens ranking and reduces spurious alerts at operational thresholds. The graph-only variant underperforms in isolation yet contributes materially within the fusion – consistent with the observed drop in calibration error and the downstream reductions in rumor half-life.

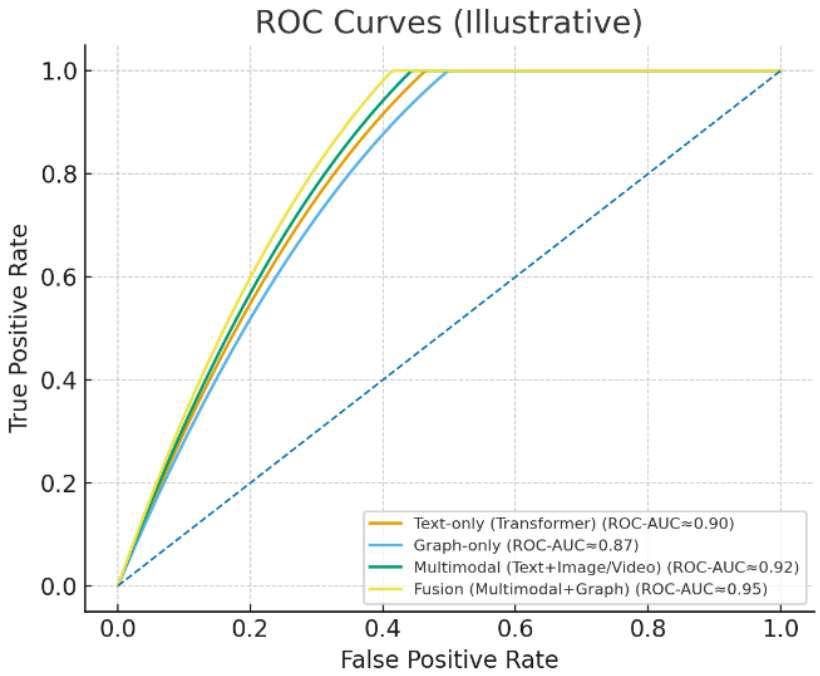

Fig 1. Receiver Operating Characteristic (ROC) curves for four model variants; diagonal denotes no-skill. Areas are parameterized to match target ROC-AUCs (Text 0.90, Graph 0.87, Multimodal 0.92, Fusion 0.95).

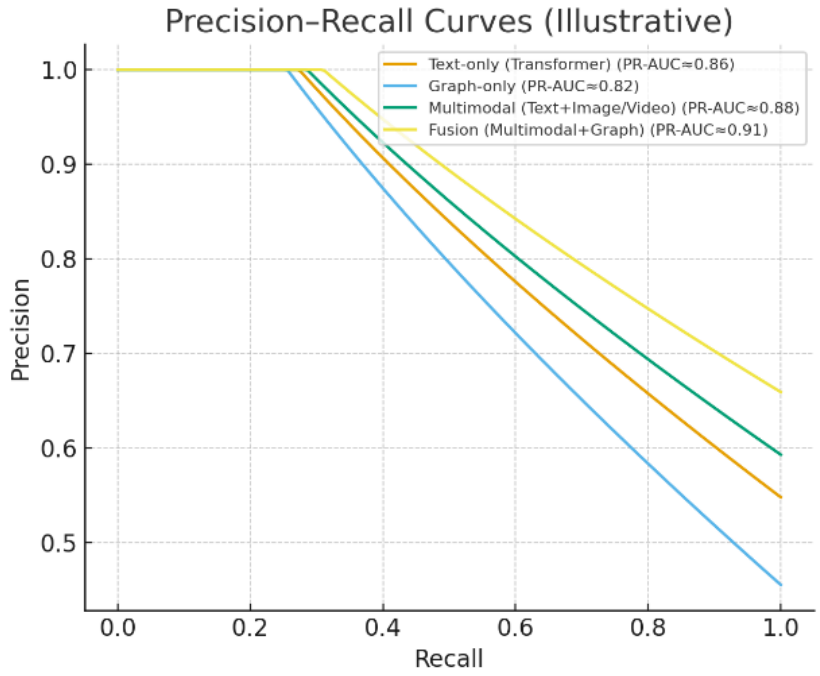

Fig 2. Precision–Recall curves with target PR-AUCs (Text 0.86, Graph 0.82, Multimodal 0.88, Fusion 0.91) under class imbalance.

4. Discussion

The comparative analyses of model performance and field outcomes indicate that fusing multimodal content with propagation structure provides a substantive and operationally meaningful advantage for enterprise disinformation defense (Saadi et al. 2025). On held-out data, the fused model not only achieved the highest F1 and PR-AUC, but also delivered the lowest calibration error, suggesting that it ranks incidents correctly and

assigns probabilities that are trustworthy for thresholding. In practical terms, this combination of discrimination and calibration translated in deployment into faster containment, shorter time-to-detection, and lower analyst workload at the same precision requirement (Żywiołek, Szymonik & Smal 2025). The joint ROC/PR perspective is important here: the ROC advantage shows more true positives for any given false-positive rate, while the PR advantage confirms that early alerts can be issued without diluting precision as coverage expands. The pattern of gains across variants clarifies where value is created. Moving from text-only to multimodal improves detection by capturing visual manipulations and cross-modal inconsistencies, but the decisive step is adding graph features that encode dissemination dynamics, community structure, and source–content relationships. In isolation, the graph model lags the text baseline, which is consistent with content still carrying strong direct signal; however, when combined, graph cues sharpen ranking and reduce spurious alerts, as evidenced by the drop in calibration error and improved early-horizon precision (García Márquez, Hameed & Jamil 2026). This complementarity aligns with the threat model: sophisticated campaigns increasingly blend plausible text with edited imagery and coordinated amplification, making single-modality defenses brittle under paraphrase and distribution shift.

Operational results reinforce the statistical findings. The observed reduction in rumor half-life and the compression of detection latency are consistent with better early precision and more reliable probabilities, which lower the cost of acting sooner. Sectoral and country heterogeneity – stronger effects in finance and healthcare and a slightly larger improvement in the UK – likely reflect differences in baseline monitoring maturity, platform mixes, and incident profiles; nevertheless, the direction of effects is stable across contexts, supporting external validity. Equally important is calibration: well-calibrated scores enable fixed, auditable thresholds tied to risk appetite, which is essential for regulated environments and post-incident accountability (Murru & Carlo 2025).

Several limitations bound interpretation. First, even with chronological splits and domain-shift tests, non-stationarity remains a moving target; the system's robustness to novel attack strategies and emergent platforms must be reassessed periodically. Second, the gold-standard labels – while showing acceptable inter-coder agreement – cannot fully represent the long tail of ambiguous or evolving narratives; residual label noise may depress absolute scores and blur small differences between models. Third, our OSINT-driven reconstruction of propagation and the mapping of enterprise logs to public items, though high-coverage, may miss private-channel dynamics, limiting insight into covert seeding. Finally, survival analyses assume reasonably stable hazard within deployment windows; abrupt exogenous shocks could alter decay dynamics in ways not fully captured by our models.

The findings nevertheless offer clear implications for practice and research. For security operations, the evidence favors deploying fused pipelines with explicit calibration and governance hooks, rather than relying on strong single-modality detectors. Playbooks should exploit calibrated scores to set tiered, business-aligned thresholds and to prioritize counter-messaging when the predicted impact channel is market- or safety-critical. For model development, future work should deepen robustness to LLM-mediated laundering and cross-lingual shift, incorporate active learning to refresh labels where uncertainty concentrates, and couple detection with causal attribution to distinguish organic virality from coordinated manipulation. Extending evaluation to include cost-sensitive metrics and budgeted alerting will further connect model behavior to analyst capacity. In sum, the

integration of content, context, and credibility is not a marginal refinement but a necessary condition for reliable, auditable, and timely enterprise response to disinformation.

5. CONCLUSION

In summary, the study demonstrates that integrating content, multimodal, and propagation signals into a single, scalable detection pipeline materially improves both the effectiveness and reliability of enterprise disinformation defense. The fused model not only raises core classification metrics and probability calibration but also yields tangible operational gains: shorter time to detection, a reduced rumor half-life, and lower analyst workload at a fixed precision threshold. Practically, these results enable fixed, auditable alert thresholds aligned with risk appetite, prioritization of incidents by predicted impact channel (market, operational, legal, reputational), and the use of early warnings to trigger rapid counter-messaging and containment. Conceptually, the findings underscore that network diffusion and source structure provide information that is complementary to content, and that well-calibrated probabilities are the crucial bridge from machine-learning metrics to defensible, time-pressured decisions in regulated environments.

The study's limitations stem primarily from non-stationarity and an adaptive adversary: even with chronological splits and domain-shift tests, robustness against novel tactics, emerging platforms, and new manipulation forms cannot be guaranteed. Despite acceptable inter-coder agreement, gold-standard labels inevitably contain noise in borderline or evolving narratives, which can depress absolute scores and blur small performance gaps. Mapping enterprise logs to public OSINT captures open-channel dynamics but may miss private or closed-group seeding. Finally, external validity, while supported by heterogeneity analyses, is bounded by the three regulatory-linguistic contexts and sector mix represented in the sample.

Future research should systematically harden robustness against LLM-mediated laundering (paraphrase, style transfer, cross-lingual translation) and topic/language drift, and incorporate active learning to refresh labels where uncertainty concentrates. Evaluation should expand to cost-sensitive metrics and budgeted alerting that tie decision thresholds to SOC capacity, and to intervention studies (e.g., A/B tests) that measure the causal impact of counter-messaging and escalation strategies. Promising directions include coupling detection with causal attribution and network forensics to distinguish organic virality from coordinated manipulation, integrating market and operational telemetry for early harm estimation, and strengthening governance with continuous calibration, drift monitoring, rationale logging, and controlled human-in-the-loop escalation. Longitudinal replications across additional jurisdictions and sectors will help assess durability of effects and refine deployment standards for enterprise disinformation defense.

REFERENCES

Angrist, J.D., Pischke, J-S., 2009, *Mostly Harmless Econometrics,* Princeton University Press.

Benjamini, Y., Hochberg, Y., 1995. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*, Journal of the Royal Statistical Society. Series B (Methodological), 57(1), 289–300. Available from: http://www.jstor.org/stable/2346101.

Chalehchaleh, R., Farahbakhsh, R., Crespi, N., 2025. *Addressing data scarcity in multilingual fake news detection: an LLM-based dataset augmentation approach*, Social Network Analysis and Mining, 15(1).

Chaudhari, D., Mhatre, K., Jadhav, B., Jakhotia, P., Mangela, M., 2026. *Securing Media Integrity: A Blockchain-Based Approach Against AI-Generated Deepfakes*, in Advanced Computing and Intelligent Technologies, eds S Das, M Paprzycki, A Ghosh & M Bianchini, Springer Nature Singapore, Singapore, 465–480.

Cherep, O., Kaliuzhna, Y., Mykhailichenko, L., Markova, S., Naumenko, Y., 2025. *Information and control systems,* Technology audit and production reserves, 2(82), 74–79.

Cox, D.R., 1972. *Regression Models and Life-Tables,* Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187–220. Available from: http://www.jstor.org/stable/2985181.

Del Moral Pérez, M.E., López-Bouzas, N., Castañeda Fernández, J., Del Bellver Moreno, M.C., 2025. *Universitarios frente a las Fake News generadas por Inteligencia Artificial: estrategias asociadas al pensamiento crítico que adoptan*, Educación XX1, 28(2), 69–121.

Demir, G., Darıcı, S., Tamer, Z., Taydaş, O., Pamučar, D., 2025. *Even I believed it! How is it possible?" The disinformation exam of journalists in the AI age: fuzzy logical approach*, Journal of Computational Social Science, 8(4).

Farhan, M., Butt, U., Sulaiman, R.B., Alraja, M., 2025. *Self-Sovereign Identities and Content Provenance: VeriTrust – A Blockchain-Based Framework for Fake News Detection*, Future Internet, 17(10), 448.

García Márquez, F.P., Hameed, A.A., Jamil, A. (eds.), 2026, *Pattern Recognition and Artificial Intelligence,* Springer Nature Switzerland, Cham.

Harris, S., Hadi, H.J., Ahmad, N., Alshara, M.A., 2025. *Multi-domain Urdu fake news detection using pre-trained ensemble model,* Scientific reports, 15(1), 8705.

Khalid, M., Mushtaq, M.F., Akram, U., Safran, M., Alfarhood, S., Ashraf, I., 2025. *Sentiment analysis for deepfake X posts using novel transfer learning based word embedding and hybrid LGR approach*, Scientific reports, 15(1), 28305.

Liang, K.Y., Zeger, S.l., 1986. *Longitudinal data analysis using generalized linear models*, Biometrika, 73(1), 13–22.

Ma, R., Wang, X., Yang, G.R., 2025. *Fighting fake news in the age of generative AI: Strategic insights from multi-stakeholder interactions, T*echnological Forecasting and Social Change, 216, 124125.

McNEMAR, Q. 1947. *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika, 12(2), 153–157.

Murru, M.F., Carlo, S., 2025. *AI and newsmaking: an exploratory investigation of discourses and practices in Italian newsrooms*, Online Media and Global Communication, 4(3), 554–569.

Nasser, M., Arshad, N.I., Ali, A., Alhussian, H., Saeed, F., Da'u, A., Nafea, I., 2025. *A systematic review of multimodal fake news detection on social media using deep learning models*, Results in Engineering, 26, 104752.

Nwaiwu, S., Jongsawat, N., Tungkasthan, A., 2025. *Decoding Disinformation: A Feature-Driven Explainable AI Approach to Multi-Domain Fake News Detection*, Applied Sciences, 15(17), 9498.

Pandey, R., Kushwaha, A.K.S., 2025. *Detecting deepfake videos: an enhanced hybrid deep learning model,* Signal, Image and Video Processing, 19(9).

Ragab, M., Ashary, E.B., Kateb, F., Hakeem, A., Mosli, R., Albogami, N.N., Nooh, S., 2025. *Classification of human-written and AI-generated sentences using a hybrid CNN-GRU model optimized by the spotted hyena algorithm*, Alexandria Engineering Journal, 126, 116–130.

Saadi, A., Belhadef, H., Guessas, A., Hafirassou, O., 2025. *Enhancing Fake News Detection with Transformer Models and Summarization*, Engineering, Technology & Applied Science Research, 15(3), 23253–23259.

Saito, K., 2026. L*egal Policy on Digital Replicas'* in Crisis or Redemption with AI and Robotics? The Dawn of a New Era, eds M.F., Silva, M.O., Tokhi, M.I.A., Ferreira, B., Malheiro, P., Guedes, P., Ferreira, M.T., Costa, Springer Nature Switzerland, Cham, 154–160.

Singh, S.K., Assi, S., Ginige, T., Mohammed, A.H., B., Wahit, F., Al-Jumeily, OBE, D., 2026. *Fake News Detection Using Explainable Artificial Intelligence* in Data Science and Emerging Technologies, eds Y.B., Wah, D., Al-Jumeily & M.W., Berry, Springer Nature Singapore, Singapore, 477–488.

Sun, L., Abraham, S., 2021. *Estimating dynamic treatment effects in event studies with heterogeneous treatment effects*, Journal of Econometrics, 225,(2), 175–199.

Zadrozny, B., Elkan, C., 07232002. *Transforming classifier scores into accurate multiclass probability estimates,* Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, eds O.R., Zaïane, R., Goebel, D., Hand, D., Keim, R., Ng., ACM, New York, NY, USA, 694–699.

Żywiołek, J., 2025a. *Creating Knowledge-Based Value for Data Security in Enterprises*, European Conference on Knowledge Management, 26(2), 1136–1145.

Żywiołek, J., 2025b. T*he Impact of Generative Artificial Intelligence on Knowledge Exchange in Supply Chains*, European Conference on Knowledge Management, 26(2), 1146–1154.

Żywiołek, J., Szymonik, A., Smal, T., 2025, *Navigating Supply Chain Turbulence,* Productivity Press, New York.