Borneo International Journal eISSN 2636-9826; Vol.8 (4); 2025; 25-34
Published by *Majmuah Enterprise*
www.majmuah.com

The Spread of Fake News on Social Media and its Implication for Public Safety

LAI WOOI SIN*, MOHAMAD FADLI ZOLKIPLI
School of Computing, College of Arts and Sciences, Universiti Utara Malaysia,
06010 Sintok, Kedah Darul Aman, Malaysia
*Email: *laiwooisin@gmail.com, m.fadli.zolkipli@uum.edu.my | Tel: *+60124212196 | Tel: +60177247779|*

Received: November 19, 2025
Accepted: November 25, 2025
Online Published: December 01, 2025

Abstract

Background: Social media platforms have become primary channels for news and crisis communication, yet their speed, scale, and engagement optimized ranking systems also enable the rapid spread of false and misleading content ("fake news"). In safety critical contexts, misinformation can distort risk perceptions, erode trust, and precipitate harmful behaviors. **Objective:** To synthesize contemporary evidence on (i) how fake news propagates across platforms; (ii) the mechanisms linking online exposure to offline public-safety harms in health emergencies and disasters; and (iii) the effectiveness and limitations of technical, design, and governance interventions. **Methods:** We conduct a narrative synthesis (2017–2025) spanning communication science, public health, information systems, and computer science. We map mechanisms along the pathway *exposure → belief → behavior → safety outcomes*, and evaluate intervention classes: transformer-based NLP, graph neural networks, multimodal/video methods, warning labels and accuracy prompts, UX friction, provenance cues, and infodemic management frameworks.

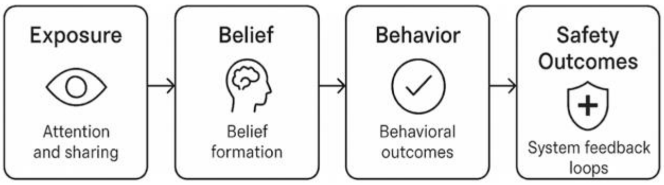

Figure 1: Pathway from exposure to harm

Results: False content benefits from novelty and affect, achieving wider and faster cascades than true content. Text-only detectors perform well in domain but degrade under domain shift; graph aware and hybrid approaches improve early detection and generalization, with emerging multimodal methods addressing video centric platforms. Field and platform studies show light-touch accuracy prompts and well-designed labels can reduce sharing of low-quality content, though effects vary by placement, specificity, and audience. Major gaps persist in (i) measuring real-world safety outcomes, (ii) robustness across topics, languages, and modalities, and (iii) transparency and data access for independent evaluation. **Conclusions:** No single solution suffices. We propose a socio technical framework integrating hybrid detection stacks, privacy safe audit pipelines, crisis aware platform design (friction, provenance, correction UX), and cross-sector coordination with media literacy and trusted-messenger strategies. A research and policy agenda is outlined to standardize evaluations, enable privacy safe data sharing, and maintain crisis playbooks that align platform incentives with public safety goals.

Keywords: misinformation; fake news; social media; public safety; risk communication

1. Introduction

1.1 Background and problem statement

Social platforms such as Facebook, X/Twitter, YouTube, TikTok, and WhatsApp have become dominant gateways for news consumption and crisis information. Ranking algorithms optimized for engagement, combined with social endorsement signals (likes, shares, comments), can unintentionally prioritize content that is novel, affect-laden, or identity-affirming traits common to misinformation. Foundational work shows that false stories travel farther and faster than true ones in social networks; while automation contributes, human psychology and social dynamics remain central drivers (Vosoughi et al., 2018), (Bovet & Makse, 2019). In safety-critical settings (e.g., pandemics, wildfires, floods), rumour cascades can misdirect attention and resources, reduce adherence to protective guidance, undermine trust in authorities, and in extreme cases precipitate harmful behaviors. From a cybersecurity perspective, coordinated influence campaigns, inauthentic amplification, and platform manipulation complicate threat detection and response. From a public-health and emergency-management perspective, the challenge is ensuring timely, trusted, and actionable guidance can compete in feeds saturated with misleading claims. The intersection of these concerns motivates an integrated analysis of risks, detection, and governance.

1.2 Objectives and research questions

This paper pursues four objectives: (i) synthesize how fake news spreads on social media; (ii) map the mechanisms linking exposure to public-safety harms in health and disaster contexts; (iii) evaluate technical, design, and governance interventions; and (iv) propose a socio-technical framework and research agenda. We address:

- RQ1: What sociotechnical factors (content, users, networks, algorithms) drive fake-news diffusion?
- RQ2: Through what pathways does exposure translate into offline risk?
- RQ3: Which interventions show credible, replicable reductions in harmful sharing or improved decision-making?
- RQ4: What governance and transparency mechanisms are needed to align platform incentives with public safety?

1.3 Scope and contributions

We review peer-reviewed literature (2017–2025) across communication science, public health, HCI, information systems, and computer science, emphasizing studies with large-scale data or causal identification. Our contributions are: (1) an integrative pathway model from exposure to safety outcomes; (2) a comparative analysis of detection and intervention evidence; (3) a pragmatic framework combining hybrid detection, crisis-aware design, and governance; and (4) an agenda for standardized evaluations and privacy-safe transparency.

2. Literature Review

2.1 Defining misinformation, disinformation, and "fake news"

Misinformation refers to false or misleading content shared without intent to deceive; disinformation denotes intentional deception, often coordinated; misinformation concerns true information used out of context to cause harm. Fake news typically indicates fabricated or heavily distorted news-like content circulating on social platforms Lazer et al., 2018), (Aïmeur, Lawani, & Dalkir, 2023). Public-health bodies conceptualize the infodemic as an overabundance of information some accurate, much not that impedes protective behaviors during emergencies (Eysenbach, 2020), (Calleja et al., 2021).

2.2 Diffusion dynamics and sociotechnical drivers

Empirical studies reveal recurring drivers: content novelty and emotion; cognitive shortcuts in fast feeds; social identity and influencer effects; network homophily and modularity; and engagement-optimised ranking (Vosoughi, Roy, & Aral, 2018), (Bovet & Makse, 2019), (Shahbazi, Rahman, & Abdullah, 2024), (Shu, Sliva, Wang, & Tang, 2017). Cross-platform flows and messaging-app forwarding accelerate cascades and complicate traceability.

2.3 Public-safety harms

Health emergencies: Exposure to vaccine misinformation lowers uptake and increases hesitancy; prompts and transparent communication attenuate these effects (Pennycook, Binnendyk, Newton, & Rand, 2022), (Calleja et al., 2021), (Altay, Berinsky, & Rand, 2025). **Disasters:** Rumours about shelters, resources, or hazard severity misdirect citizens and responders (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Ilyas et al., 2025). **Trust and compliance:** Persistent misinformation erodes institutional trust and lowers compliance with advisories, undermining collective action (Shahbazi, Rahman, & Abdullah, 2024), (Altay, Berinsky, & Rand, 2025), (Pang et al., 2023).

2.4 Detection technologies

Text-centric detectors (BERT/RoBERTa) perform well in-domain but suffer under distribution shift (Hartwig, Biselli, Schneider, & Reuter, 2024). Graph-based methods (GNNs) that model propagation and user item networks improve early detection and cross-topic generalization Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024); graph refinement (e.g., DECOR) denoises edges and helps at scale (Wu & Hooi, 2023). Multimodal approaches are increasingly vital for video-first platforms (Hilberts & Bernard, 2025), (Muenster et al., 2024).

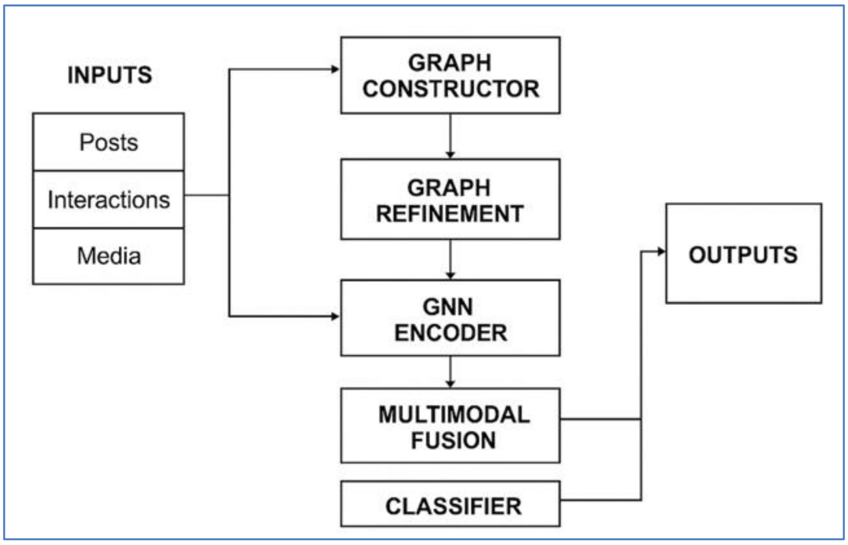

Figure 2: Hybrid fake news detection architecture

Table 1: Detection approaches capabilities and limitations

Approach	Signals	Strengths	Limitations	Suitability
Transformer (text)	Lexical/semantic	Strong in-domain accuracy; fast inference	Domain shift; language/topic sensitivity	Articles, long posts
Propagation trees	Cascade structure	Early signals; platform-agnostic	Needs interaction data; privacy constraints	Rapidly spreading rumours
GNN on user item graph	Topology, communities	Robust to wording; captures coordination	Cold-start for new nodes; data access	Cross-topic generalization
Multimodal	Frames, ASR, OCR	Handles short video platforms	Heavy compute; eval standards lag	TikTok, Reels, Shorts
Graph refinement	Denoised edges	Improves GNN quality, scalability	Complexity, tuning	High-volume events

2.5 Interventions and governance

Behavioural nudges and labels: Accuracy prompts reliably reduce intentions to share low-quality content; warning labels help but effects depend on placement and specificity (Pennycook, Binnendyk, Newton, & Rand, 2022), (Ling, Gummadi, & Zannettou, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024). **Friction and provenance:** Re-share friction (e.g., "read before share") and provenance metadata can slow diffusion; rigorous at-scale evidence is emerging. **Institutional capacity:** WHO-aligned infodemic-management frameworks emphasize listening, targeted engagement, and resilience building (Eysenbach, 2020)–(Calleja et al., 2021), (Pang et al., 2023). **Transparency:** Researcher access to ranking and enforcement data remains limited, constraining policy learning (Lazer et al., 2018).

3. Methodology

We synthesize evidence along the pathway exposure → belief → behavior → safety outcomes, grading studies by design strength and external validity. We map intervention points to that pathway and compile an evidence matrix across outcomes (exposure reduction, sharing reduction, belief accuracy, compliance, crisis operations).

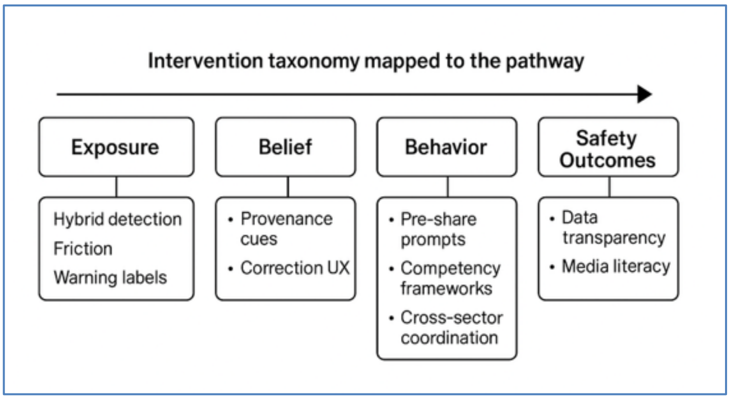

Figure 3:Pathways from exposure to harm

Table 2: Intervention evidence map

Intervention class	**Representative studies**	**Primary mechanism**	**Exposure↓**	**Sharing ↓**	**Belief accuracy↑**	**Behavioral compliance ↑**	**Notes**
Accuracy prompts	Pennycook et al. 2022 (Pennycook, Binnendyk, Newton, & Rand, 2022)	Salient accuracy goal	△	⊠⊠	⊠	△	Robust across topics; small-to-moderate effects
Warning labels	ICWSM/CHI TikTok studies (Ling, Gummadi, & Zannettou, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024)	Risk cueing, salience	△	⊠	△ ⊠	△	Depends on placement/specificity; possible reactance
Re-share friction	Platform UX reports	Costly sharing	⊠	⊠	△	△	Strong theory; more causal field evidence needed
Provenance cues	Platform/standards pilots	Source transparency	△	△	⊠	△	Most evidence in lab/observational contexts
Hybrid detectors	GNN+Transformer (Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024), (Wu & Hooi, 2023)	Early cascade flagging	⊠	⊠	—	—	Systems metrics strong; link to safety outcomes scarce
Infodemic mgmt.	WHO frameworks (Eysenbach,	Listening + targeted	△	△	⊠	⊠	Agency capacity central; measurement

Borneo International Journal eISSN 2636-9826; Vol.8 (4); 2025; 25-34
Published by *Majmuah Enterprise*
www.majmuah.com

	2020)–(Calleja et al., 2021), (Pang et al., 2023)	comms					varies

Legend

✔✔	Consistent medium+ effect
✔	Positive effect
Δ	Mixed/limited evidence
—	Not directly measured.

4. Discussion & Analysis

4.1 Pathways from exposure to harm

Attention advantages for novel/affect-rich falsehoods amplify reach; accuracy prompts restore deliberation and reduce sharing (Vosoughi, Roy, & Aral, 2018), (Pennycook, Binnendyk, Newton, & Rand, 2022). In health crises and disasters, misinformation can degrade compliance and misdirect resources (Eysenbach, 2020), (Calleja et al., 2021), (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Ilyas et al., 2025). Engagement-weighted ranking and community clustering reinforce exposure (Vosoughi, Roy, & Aral, 2018), (Shahbazi, Rahman, & Abdullah, 2024), (Shu, Sliva, Wang, & Tang, 2017).

4.2 Strengths and gaps in detection

Strengths, Hybrid models outperform text-only baselines; graph refinement and early-tree features aid fast events (Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024), (Wu & Hooi, 2023). **Gaps**, cross-domain robustness, adversarial adaptation, measurement of offline safety outcomes, and transparency for audits (Lazer et al., 2018), (Shu, Sliva, Wang, & Tang, 2017), (Hartwig, Biselli, Schneider, & Reuter, 2024)..

4.3 Interventions that show promise

Low-friction accuracy prompts, well-designed labels, forward-limit friction, provenance cues, and organizational infodemic capacity are the most credible approaches; each needs rigorous, transparent evaluation at platform scale (Pennycook, Binnendyk, Newton, & Rand, 2022), (Calleja et al., 2021), (Pang et al., 2023), (Ling, Gummadi, & Zannettou, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024).

4.4 Ethical, legal, and societal considerations

Balance safety with expression; avoid disproportionate harms; ensure contestability and explainability of labels/enforcement; provide privacy-safe transparency for independent audits (Lazer et al., 2018), (Shahbazi, Rahman, & Abdullah, 2024), (Shu, Sliva, Wang, & Tang, 2017).

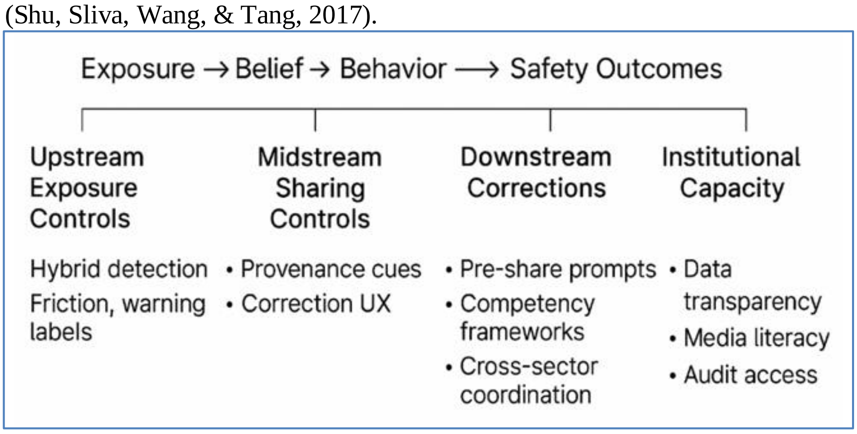

Figure 4: Intervention taxonomy mapped to the pathway

This figure, **Upstream Exposure Controls** act before users form beliefs by reducing how much low-quality content is surfaced (e.g., hybrid detection flags fast-growing cascades; friction and warning labels add cost/salience to re-sharing). Primary metrics: impressions, reach, reproduction rate (R). **Midstream Sharing Controls** Target Belief formation, providing provenance cues and correction friendly UX that raise accuracy salience at the decision moment. Primary metrics: share through rate, click through to sources, belief accuracy. **Downstream Corrections** intervene after content spreads to shift Behavior elevating verified fact-checks, enabling user retractions without penalty, and deploying trusted messengers. Primary metrics: reduction in onward shares after correction, correction visibility, compliance with guidance. **Institutional Capacity** underpins Safety Outcomes via data transparency and audit access for independent evaluation, competency frameworks for agencies, and cross sector coordination during crises. Primary metrics: time to correct, coverage in affected locales/languages, operational KPIs (e.g., evacuation compliance, vaccination uptake).

How to use the taxonomy. Select interventions by crisis phase (pre-event preparedness vs. surge), platform modality (text vs. short video), and constraints (data access/privacy). Combine one or more controls in each band for layered defense, and pre-register outcome metrics to enable causal evaluation.

Table 3: Governance and transparency checklist

Domain	Minimal standard	Better	Best
Data access	Researcher API with rate limits	Privacy-safe impression/exposure logs	Clean rooms + audit trails for interventions
Ranking transparency	High-level docs	Auditable disclosures of ranking inputs	Third-party audited impact assessments
Labelling policy	Static policy page	Versioned changelog + rationales	Public policy API + dispute metrics
Crisis protocols	Ad-hoc comms	Written surge playbooks	Cross-sector exercises & after-action reports
Measurement	Posterior metrics (CTR)	Pre-registered outcome metrics	Causal evaluations tied to safety outcomes

Summarizes a governance and transparency maturity ladder across five domains data access, ranking transparency, labelling policy, crisis protocols, and measurement. At the minimal level, platforms offer only a basic researcher API, high-level descriptions of ranking, static policy pages, ad-hoc crisis communications, and engagement-based metrics. Moving to better practice adds privacy-preserving impression/exposure logs, auditable disclosures of ranking inputs, versioned policy changelogs with rationales, written surge playbooks, and pre-registered outcome measures. Best-in-class practice includes secure data clean rooms with intervention audit trails, third-party impact assessments of recommender systems, a public policy API with dispute/appeal metrics, regular cross-sector exercises with after-action reports, and causal evaluations tied to real-world safety outcomes such as evacuation compliance or vaccine uptake. The checklist is intended for platforms, public agencies, and researchers to shift at least one step rightward in each domain and to track KPIs such as time-to-correct, post-label share-through, appeal turnaround, and coverage of executed surge playbooks.

5. Proposed Solution

An effective response to fake news and its public safety impacts requires a layered, socio-technical strategy that aligns platform engineering, product design, institutional governance, and public capacity building. On the technical front, platforms should deploy **hybrid detection stacks** that blend transformer-based text encoders with graph refinement neural networks and propagation features, while integrating multimodal fusion for short-video and image-heavy contexts (Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024), (Wu & Hooi, 2023). Such systems ought to be evaluated for **robustness under distribution shift** using multilingual, cross-topic, and cross platform benchmarks, with explicit early detection and open set conditions; beyond raw accuracy, models should report **calibration and uncertainty** so that triage systems can route high risk cases to human review (Hartwig, Biselli, Schneider, & Reuter, 2024). **Privacy preserving audit pipelines** such as secure data clean rooms and logged intervention traces are essential to support independent red-teaming and to link exposures and enforcement actions to downstream outcomes, a need repeatedly emphasized in governance analyses of the misinformation ecosystem (Lazer et al., 2018).

Product and UX design should complement detection rather than attempt to replace it. **Pre share accuracy prompts**, **read before share interstitials**, and **warning labels** can nudge deliberation at the moment users decide to repost or interact; when carefully designed, these interventions reduce sharing of low-quality content across topics and platforms, though effects depend on placement and specificity (Pennycook, Binnendyk, Newton, & Rand, 2022), (Ling, Gummadi, & Zannettou, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024). **Provenance signals** and source context help users appraise credibility without leaving the feed, and **correction experiences** should make it simple to retract or amend a post without disproportionate penalties, with verified corrections elevated in the same contexts where the original claims spread (aligning with WHO informed infodemic practices) (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023). During crisis surges, platforms should activate **temporary ranking policies** that prioritize authoritative guidance, dampen rapid forwarding in messaging channels, and localize information to affected languages and geographies; these surge rules should be codified in **incident playbooks** and tested during exercises (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Pang et al., 2023), (Ilyas et al., 2025).

Governance and policy provide the scaffolding that ensures interventions are transparent, accountable, and optimized for public safety rather than only for engagement. At a minimum, platforms should provide privacy-safe **researcher access** to impression and exposure data, publish auditable documentation of ranking inputs, and maintain **versioned changelogs** that explain labelling and enforcement policies (Lazer et al., 2018). More mature regimes include **third party impact assessments** of recommender systems, public **policy APIs** that expose dispute and appeal metrics, and **cross sector protocols** that enable rapid coordination among platforms, health and emergency agencies, NGOs, and fact-checkers (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023). Standard reporting templates that specify intervention scope, triggers, and intended outcomes can improve comparability and external evaluation across events and platforms (Lazer et al., 2018).

Building **public resilience** is the durable complement to system level controls. Scalable **media-literacy programs** that emphasize accuracy goals, lateral reading, and source triangulation can improve baseline resistance to falsehoods, particularly when paired with **trusted messengers'** clinicians, local officials, educators, and community leaders who translate corrections into culturally resonant language (Pennycook, Binnendyk, Newton, & Rand, 2022), (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023). Institutions should invest in **infodemic-management competencies**: continuous listening, message testing, multilingual outreach, and rapid correction protocols that integrate community partners (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023). Because resources are finite, agencies should pre-register target populations, channels, and measurable objectives so that interventions can be prioritized where marginal safety gains are highest (Ilyas et al., 2025).

Implementation should be guided by a **common measurement framework** that connects online actions to offline outcomes. Each intervention ought to specify leading indicators (exposures to low quality content, share through rate, time to correct, correction visibility) and, where feasible, **causal links to safety metrics** such as vaccination uptake, evacuation compliance, emergency room load, or response team efficiency (Pennycook, Binnendyk, Newton, & Rand, 2022), (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Ilyas et al., 2025). Programs should commit to **iterative evaluation** A/B tests during steady state, quasi-experiments during crises, and after-action reviews that refine playbooks and to publishing aggregate findings that advance collective learning while protecting user privacy (Lazer et al., 2018). By combining robust technical safeguards, friction and provenance in UX, transparent governance with independent evaluation, and sustained public-facing capacity, platforms and agencies can meaningfully reduce the real-world risks posed by fake news while preserving open, pluralistic online discourse.

6. Conclusion and Future Work

This paper synthesized the current state of knowledge on the spread of fake news across social media and its implications for public safety. We integrated evidence from communication science, public health, HCI, and machine learning to articulate a pathway linking exposure to false content with belief formation, behavioural responses, and downstream safety outcomes. The review showed that novelty, emotion, and identity laden content exploits engagement driven ranking and socially clustered networks to secure attention and diffuse rapidly (Vosoughi, Roy, & Aral, 2018), (Bovet & Makse, 2019). While text only detection can succeed within narrow domains, more robust performance in the wild increasingly depends on graph aware and multimodal approaches that capture propagation context and short-video dynamics (Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024), (Wu & Hooi, 2023), with persistent challenges from distribution shift and adversarial adaptation (Hartwig, Biselli, Schneider, & Reuter, 2024). Field evidence indicates that **low friction prompts** and **well-designed warning labels** can reduce sharing of low-quality content, yet their effects are sensitive to placement, specificity, and audience characteristics (Pennycook, Binnendyk, Newton, & Rand, 2022), (Ling, Gummadi, & Zannettou, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024). These findings motivate a socio technical agenda that couples technical safeguards and product design with governance mechanisms and public-facing capacity (Lazer et al., 2018), (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023).

A central conclusion is that mitigation should be measured not only by platform internal metrics (e.g., click through rate or raw classifier accuracy) but by their contribution to **public safety outcomes**, such as improved adherence to protective guidance, faster rumour correction during disasters, and reductions in resource misallocation (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Ilyas et al., 2025). Achieving this shift requires privacy safe **transparency** so independent researchers can evaluate interventions, as well as institutional competencies within public agencies to listen, engage, and communicate effectively across languages and modalities (Lazer et al., 2018), (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023). We also underscore several constraints: label scarcity and topic drift undermine generalization; platform policies change over time in ways that complicate evaluation; and heterogeneous local contexts limit the portability of one size fits all interventions (Lazer et al., 2018), (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025).

Future work should prioritize **causal identification** of intervention effects on offline safety. This includes randomized field experiments and quasi experimental designs that link exposure, labelling, and friction to concrete outcomes such as vaccination uptake, evacuation compliance, emergency-room load, or responder routing efficiency (Pennycook, Binnendyk, Newton, & Rand, 2022), (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Ilyas et al., 2025). To make these studies feasible while preserving privacy, platforms should expand secure **data access mechanisms** (e.g., clean rooms with verified logging of impressions, labels, and enforcement actions) and publish pre-registered protocols and outcome definitions (Lazer et al., 2018). Complementary audit studies can probe for heterogeneous effects by community, language, and device type to ensure that gains do not mask harms among specific populations (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023).

On the technical front, priorities include rigorous **robustness evaluation** across domains (health, disasters, elections), languages (including low resource languages and dialects), and modalities (text, image, audio, short video) (Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024), (Wu & Hooi, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024). Research should examine **open set** and **early detection** settings where ground truth is sparse, and delay costs are high; **calibration and uncertainty estimation** should be first class metrics so that models defer gracefully to human review under ambiguity (Hartwig, Biselli, Schneider, & Reuter, 2024). Graph refinement and propagation aware models merit further study for detecting coordinated networks and cross-platform cascades (Wu & Hooi, 2023), while multimodal fusion must handle compression artifacts, OCR/ASR noise, and adversarial media manipulations (Hilberts & Bernard, 2025), (Muenster et al., 2024). A related thread is **content provenance** standardizing cryptographic signatures and metadata for images and video and understanding how provenance cues interact with user trust and sharing behavior in practice (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023).

Design research should explore how **prompts, labels, provenance, and correction experiences** perform under realistic attention constraints. Key questions include the optimal timing and specificity of prompts, the trade-offs between salience and reactance for labels, and the design of non-penalizing retraction flows that encourage self-correction without amplifying false claims (Pennycook, Binnendyk, Newton, & Rand, 2022), (Ling, Gummadi, & Zannettou, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024). Messaging app environments warrant special attention: forwarding limits, small group dynamics, and end to end encryption require distinct approaches for monitoring and intervention that respect privacy while enabling public interest measurement at aggregate levels (Lazer et al., 2018), (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023).

Governance and policy research should develop **standardized reporting** for interventions (scope, triggers, intended outcomes), common data schemas for exposure and label events, and **playbook templates** for crisis surge conditions (Lazer et al., 2018), (Eysenbach, 2020), (Calleja et al., 2021), (Pang et al., 2023). Cross sector exercises that involve platforms, emergency agencies, health authorities, NGOs, and community leaders can test coordination mechanisms before crises and establish benchmarks for response times, language coverage, and the visibility of authoritative corrections (Erokhin & Komendantova, 2024), (Hilberts & Bernard, 2025), (Ilyas et al., 2025). Ethical frameworks should codify safeguards for freedom of expression, due process, and equity, including routine disparate impact assessments of ranking and enforcement policies and transparent appeal channels with published resolution metrics (Lazer et al., 2018). Finally, advancing this agenda requires an **open science posture**: sharable benchmark suites that reflect realistic, multilingual, and multimodal conditions; reproducible pipelines and evaluation code; and publication norms that encourage the release of aggregate effect sizes, confidence intervals, and negative results (Phan, Hussain, & Nguyen, 2023), (Malik & Bakhtyar, 2024), (Wu & Hooi, 2023), (Hartwig, Biselli, Schneider, & Reuter, 2024). By aligning technical innovation with careful measurement, transparent governance, and community centered practice, the research and practitioner communities can move beyond narrow accuracy gains toward interventions that measurably improve public safety during health emergencies, disasters, and other high stakes events.

Acknowledgment

The author gratefully acknowledges the mentorship of Associate Prof. Ts Dr. Mohamad Fadli Bin Zolkipli, whose guidance refined the research design and argumentation. The author also thanks the Cybersecurity in Social Media teaching team for thoughtful feedback on interim submissions and to classmates for insightful discussion that strengthened this work. Appreciation is extended to fact-checking organizations and open-access repositories whose datasets and reports enable rigorous study of misinformation and public safety.

Funding/Conflicts: No external funding was received. The author declares no conflicts of interest.

Ethics and Data Use: This study synthesizes previously published, publicly available research; no new human data were collected. When recommending interventions, the paper emphasizes privacy-preserving measurement, transparency, and accountability.

References

Altay, S., Berinsky, Z. A., & Rand, D. G. (2025). Following news on social media boosts knowledge, belief accuracy, and trust. *Nature Human Behaviour*.

Aïmeur, E., Lawani, G., & Dalkir, S. (2023). Fake news, disinformation and misinformation in social media: A review and classification. *Social Network Analysis and Mining, 13*.

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 U.S. presidential election. *Nature Communications, 10*(1).

Calleja, N., et al. (2021). A public health research agenda for managing infodemics. *Health Security, 19*(1), 12–40.

Erokhin, D., & Komendantova, N. (2024). Social media data for disaster risk management and research. *International Journal of Disaster Risk Reduction*.

Eysenbach, G. (2020). How to fight an infodemic: The four pillars of infodemic management. *Journal of Medical Internet Research, 22*(6), e21820.

Farajtabar, M., et al. (2017). Fake news mitigation via point process–based intervention. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Hartwig, K., Biselli, J., Schneider, S., & Reuter, C. (2024). Assessing an indicator-based intervention to combat misinformation on TikTok. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024)*.

Hilberts, S., & Bernard, H. R. (2025). The impact of misinformation on social media in the context of natural disasters: Narrative review. *JMIR Infodemiology, 5*(1).

Ilyas, B., et al. (2025). Social media-based sentiment analysis for disaster risk management: A systematic review. *International Journal of Disaster Risk Reduction*.

Kuntur, S., Reddy, R., & Rao, C. (2024). Comparative analysis of GNNs and Transformers for fake news detection. *Electronics, 13*(23).

Lazer, D. M. J., et al. (2018). The science of fake news. *Science, 359*(6380), 1094–1096.

Ling, C., Gummadi, K. P., & Zannettou, S. (2023). “Learn the facts about COVID-19”: Analyzing the use of warning labels on TikTok videos. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Malik, A., & Bakhtyar, M. (2024). Ensemble graph neural networks for fake news detection. *Array, 14*.

Muenster, R. M., et al. (2024). Alternative health and conventional medicine discourse on TikTok: Implications for misinformation. *Journal of Medical Internet Research*.

Pang, M. D., et al. (2023). Global overview and insights on infodemiology and infodemic management. *JMIR Infodemiology, 3*(1).

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2022). Accuracy prompts are a replicable and generalizable approach to reducing the spread of misinformation. *Nature Communications, 13*.

Phan, H. T., Hussain, N., & Nguyen, Q. V. H. (2023). Fake news detection: A survey of graph neural network methods. *Applied Soft Computing, 140*, 110155.

Rubinelli, S., et al. (2022). WHO competency framework for health authorities and institutions to manage infodemics. *Human Resources for Health, 20*, 35.

Shahbazi, M., Rahman, A. R., & Abdullah, R. M. O. (2024). Social media trust: Fighting misinformation in the time of crisis. *Information & Management*.

Shu, K., Sliva, A., Wang, S., & Tang, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, *19*(1), 22–36.

Tangcharoensathien, V., et al. (2020). Framework for managing the COVID-19 infodemic. *Journal of Medical Internet Research, 22*(6), e19659.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

Wang, W. Y. (2017). “Liar, Liar Pants on Fire”: A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL).*

World Health Organization. (2025). *Infodemic—Key concepts and management*.

Wu, J., & Hooi, B. (2023). DECOR: Degree-corrected social graph refinement for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).*

Click or tap here to enter text.