International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 6, Nov – Dec 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

A Reliability-Aware Framework for Detecting AI-Generated and Misleading Emergency Messages Using Multi-Signal Evidence Analysis

Maseeha Bhanu, Amulya D R, G V Abishek Bhargav, Ashwini C J, Suhas S L

Dept. of CSE (AI&ML), Malnad College of Engineering, Hassan, India

mb@mcehassan.ac.in, amulyavdev@gmail.com, gvabhishekbhargav14@gmail.com, Ashwinicj256@gmail.com, Suhasgowda3939@gmail.com

Abstract

Emergency communication channels are increasingly flooded with messages that are machine-generated, outdated, or deliberately misleading, which delays response and erodes public trust. Existing detectors usually depend on a single signal, such as writing style or propagation pattern, and rarely explain why a message is flagged. This paper proposes Adaptive Emergency Message Reliability Assessment (AEMRA), a framework that fuses five signals: AI-generation likelihood, source credibility, spatial consistency, temporal consistency, and retrieved evidence support. A context-aware gating network assigns signal weights and renormalizes over the signals that are actually available, so the method keeps working when a geotag or source profile is missing. AEMRA outputs a reliability score, a risk-priority rank for moderators, and a per-signal explanation. We evaluate AEMRA on [datasets] against [baselines] using [metrics]. Results show [main finding], and ablation shows [finding on adaptive weighting and missing signals].

Index Terms—*emergency messages, misinformation detection, AI-generated text, source credibility, evidence retrieval, explainable verification, crisis informatics.*

I. Introduction

During floods, earthquakes, and public-safety incidents, social media and messaging platforms become primary channels for reports and warnings. They also carry rumors, recycled content, and, increasingly, fluent text produced by large language models. A fabricated evacuation order or a false shelter location can send people into danger and divert limited responders.

Three trends make the problem harder than ordinary misinformation detection. First, generative models make plausible emergency text cheap to produce at scale. Second, emergency messages are short, time-critical, and often lack metadata such as geotags or verified profiles. Third, moderators need to know not only whether a message is unreliable but which unreliable message is most dangerous right now, and why.

Most prior systems address one aspect: text classification, propagation analysis, or fact-checking against retrieved sources. Few combine several kinds of evidence, handle absent signals gracefully, rank by potential harm, or expose the reasoning behind a verdict.

This paper makes the following contributions:

1. AEMRA, a five-signal framework that jointly estimates AI-generation likelihood, source credibility, spatial consistency, temporal consistency, and evidence support for an emergency message.
2. A context-adaptive fusion rule that learns signal weights from message context and renormalizes over available signals.
3. A risk-priority score that combines unreliability with severity, reach, and urgency to order moderator queues.
4. A per-signal explanation that reports each signal's contribution and the strongest supporting or contradicting evidence passage.
5. An evaluation protocol with baselines, ablations, calibration, ranking, and missing-signal tests, plus a documented method for building an AI-generated emergencymessage test set.

The remainder of the paper reviews related work, states the problem, presents the method, architecture, and algorithm, describes the implementation and experiments, and concludes.

II. Related Work

A. Crisis informatics and rumor detection

Crisis datasets such as CrisisMMD and HumAID [1], [2] support classification of informative and humanitarian content in disaster posts. Rumor and misinformation studies show that false claims spread differently from true ones [3] and that credibility can be estimated from content and user features [4]. These works motivate source and propagation features, but they mainly treat each as a separate classifier.

B. Detecting machine-generated text

Detectors range from fine-tuned transformer classifiers [5] to zero-shot methods based on probability curvature [6]. Their accuracy drops under paraphrasing, domain shift, and short inputs, all common in emergency messages. A detector score is therefore better treated as one uncertain signal than as a verdict.

C. Retrieval and evidence-based verification

Fact verification benchmarks such as FEVER [7] and natural language inference models [8] check claims against evidence

 ©IJSRED: All Rights are Reserved

passages. Retrieval-augmented generation [9] provides a way to bring in current sources. In emergencies, evidence must come from fast, authoritative feeds, and retrieval can return nothing, which a verifier must handle explicitly.

D. Multi-signal and adaptive fusion

Gated and mixture-of-experts models [10] learn inputdependent weights over components. Calibration work [11] shows that neural scores are often overconfident, which matters when scores drive moderation priority.

E. Recent emergency-specific work

Recent studies combine large and small language models, retrieval augmentation, and multi-agent credibility scoring for emergency rumor detection and fact-checking. These approaches strengthen individual components such as the language model or the retriever. AEMRA instead concentrates on fusing heterogeneous signals under missing-data conditions, ranking by potential harm, and explaining each verdict.

III. Existing System

Deployed and published pipelines typically fall into four families: text-only classifiers, source or account reputation scoring, propagation and network analysis, and retrieval-based fact-checking. Each is strong on one signal and blind to the others. Table I summarizes these families together with AI-text detectors, which are now often added as a fifth check.

TABLE I. Families of Existing Systems and Their Limitations

Family	Typical signal	Strength	Limitation for emergencies
Text-only classifiers	Wording, style, topic	Fast; needs no metadata	Fooled by fluent generated text; no grounding in facts
Source reputation scoring	Account age, verification, history	Cheap, stable	New or anonymous witnesses are penalized; accounts can be hijacked
Propagation and network analysis	Spread pattern, retweet graph	Catches coordinated campaigns	Needs time to accumulate; too slow for the first minutes
Retrieval-based fact-checking	Match against retrieved evidence	Grounded in sources	Fails when no official source exists yet; adds latency
AI-text detectors	Machinegeneration likelihood	Flags synthetic content	Unreliable on short, paraphrased text; says nothing about truth

IV. Research Gap

The gap AEMRA targets has three parts:

1. Signals are combined with fixed or hand-set weights that ignore disaster type, language, or channel.
2. Missing signals, such as no geotag or no source history, are handled by imputation or silently lowered scores rather than principled renormalization.
3. Systems output a label or score without ranking by potential harm and without a per-signal explanation a moderator can act on.

Table II compares typical existing approaches with AEMRA on these capabilities.

TABLE II. Capability Comparison Between Existing Approaches and AEMRA

Capability	Typical existing approach	AEMRA
Signals combined	One, or a few with fixed weights	Five, with context-adaptive weights
Missing signals	Imputed or ignored silently	Renormalized over available signals
Ranking	Binary or probability label	Risk priority using severity, reach, urgency
Explanation	None or a single feature list	Per-signal contributions plus evidence passage
Calibration	Rarely reported	Trained and evaluated with ECE

V. Problem Statement

Given an emergency message m with text x, source s, claimed location l, timestamp t, and an incomplete set of available signals, estimate a reliability score R(m) in [0, 1] that is well calibrated, rank messages by risk priority so that the most dangerous unreliable messages surface first, and return an explanation of each signal's contribution.

VI. Proposed Methodology: AEMRA

For each message m, AEMRA computes five signals normalized to [0, 1], where higher means more reliable.

A. Signal estimation

1) AI-generation likelihood. An ensemble of a fine-tuned transformer classifier and perplexity-style features gives pAI(x), the probability that the text is machine-generated. The signal is: $z1 = 1 - pAI(x)$

2) Source credibility. Features $\varphi(s)$ include verified status, authority registry match, account age, and historical accuracy. A learned linear layer with a sigmoid gives: $z2 = \sigma(w^T \varphi(s))$

3) Spatial consistency. The distance d between the claimed location l and a reference location lref (from geotags, extracted place entities, or an official incident feed) is mapped by a decay with scale τd:

$$z3 = \exp(-d(l, lref) / \tau d)$$

4) Temporal consistency. The gap between the message time t and the event time tref from the authoritative feed is mapped with scale τt:

$$z4 = \exp(-|t - tref| / \tau t)$$

5) Evidence support. Passages ek are retrieved from official sources such as agency bulletins and weather services. An NLI model returns entailment and contradiction probabilities, which are averaged with retrieval weights αk and rescaled to [0, 1]:

$$q = \Sigma k\, \alpha k\, (Pent(m, ek) - Pcon(m, ek)) / \Sigma k\, \alpha k, \quad z5 = (q + 1) / 2$$

B. Adaptive fusion with missing signals

Let ai be 1 if signal i is available and 0 otherwise (for example, a3 = 0 when there is no usable location; a1 = 0 only if the detector returns no score). A gating network over the

 ©IJSRED: All Rights are Reserved

context vector c, which encodes disaster type, language, channel, and the availability mask, produces weights: w = softmax(gθ(c))

The reliability score renormalizes over available signals:

$$R(m) = \Sigma i=1..5 \text{ ai wi zi} / \Sigma i=1..5 \text{ ai wi}$$

If no signal is available, R(m) is undefined and the message is routed to manual review. *C. Risk prioritization*

Severity Sev, reach Reach, and urgency Urg are each normalized to [0, 1]. Severity comes from the incident type, reach from shares or audience size, and urgency from the time left before the claim could cause harm. The priority is:

$$\Pi(m) = (1 - R(m)) \cdot Sev(m) \cdot Reach(m) \cdot Urg(m)$$

D. Explanation

Each signal's contribution is: ci = ai wi zi / Σj aj wj

The explanation lists the ci values, flags the lowest as the main concern, and shows the top supporting or contradicting evidence passage.

E. Training

Only the gating parameters θ and the credibility weights w are trained, with the detector and NLI model frozen or finetuned separately. The loss is binary cross-entropy on R(m) plus a calibration penalty:

$$\mathscr{L} = BCE(R(m), y) + \lambda \cdot ECEsoft$$

To make the method robust to missing signals, random signals are masked during training so the gate learns to rely on what is present.

VII. System Architecture

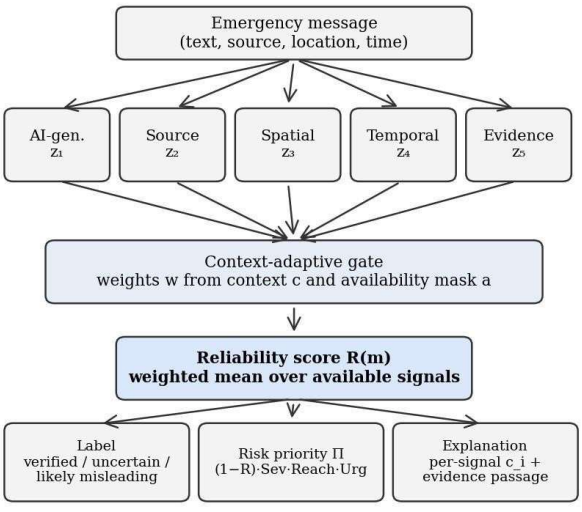


Fig. 1. Architecture of AEMRA. Each signal is scored independently. The gate weights only the signals that are present, and the fused score R(m) drives the label, the priority rank, and the explanation.

VIII. Proposed Algorithm

Algorithm 1 scores one message and returns a label, a priority, and an explanation.

ALGORITHM 1. AEMRA Scoring

Input: message m = (x, s, l, t), parameters θ, scales τd and τt, thresholds (Thi, Tlo)
Output: label, priority Π, explanation E

(x, entities, l, t) ← Preprocess(m)
z1 ← 1 − pAI(x); a1 ← 1 if the detector returns a score, else 0
z2 ← σ(w^T φ(s)); a2 ← 1 if a source profile exists, else 0
z3 ← exp(−d(l, lref) / τd); a3 ← 1 if l and lref exist, else 0
z4 ← exp(−|t − tref| / τt); a4 ← 1 if t and tref exist, else 0
Evid ← Retrieve(x, official sources)
if Evid is empty: a5 ← 0; else z5 ← (Aggregate(NLI(x, Evid)) + 1) / 2, a5 ← 1
c ← BuildContext(disaster type, language, channel, a)
w ← Softmax(gθ(c))
if Σ ai wi = 0: return ('manual review', –, –)
R ← Σ ai wi zi / Σ ai wi
Π ← (1 − R) · Sev · Reach · Urg
label ← 'verified' if R ≥ Thi; 'likely misleading' if R < Tlo; else 'uncertain'
E ← {ci = ai wi zi / Σ aj wj} + best supporting/contradicting passage
return (label, Π, E)

Complexity. Cost per message is dominated by detector and NLI inference plus retrieval. Fusion itself is O(1) in the number of signals (five). Latency is measured in Section X.

IX. Implementation

The AEMRA prototype is implemented in Python. Table III summarizes its components.

TABLE III. AEMRA Implementation Components

Component	Suggested implementation
AI-text detector	Fine-tuned RoBERTa-base classifier plus perplexity features from a small causal LM
Source credibility	Logistic layer over account features, with an authority registry (government, meteorological, and disaster-agency handles)
Location and time	Named-entity recognition for places and dates, geocoding to coordinates, haversine distance to incident feed location
Evidence retrieval	BM25 plus dense retriever over a cached corpus of official bulletins
Evidence scoring	Pretrained NLI model (MNLI-style) over message and passage pairs
Gating network	Two-layer MLP over context embedding, softmax output over five signals, PyTorch
Serving	FastAPI service, with a moderator view showing label, priority, and explanation

Data pipeline. Messages are cleaned (URL and mention normalization), language-tagged, and enriched with the availability mask. Official bulletins are cached with timestamps so retrieval is reproducible. Random masks over signals 2 to 5 are applied during training.

Reproducibility. We fix random seeds, log model versions and dataset snapshots, and release configuration files. We also report hardware and per-message latency.

X. Experimental Setup and Results

A. Datasets

Real emergency content is drawn from public crisis corpora, namely HumAID and CrisisMMD [1], [2]. An AI-generated and adversarial subset is constructed by prompting several large

 ©IJSRED: All Rights are Reserved

language models to write fabricated emergency posts, including false orders, wrong locations, and stale events. The prompts, models, and filtering rules are documented in the released configuration. Data are split by event rather than by message to avoid leakage. The final dataset contains [N] real, [N] AIgenerated, and [N] misleading messages.

B. Baselines

Table IV lists the baselines.

TABLE IV. BASELINE METHODS

ID	Baseline	Purpose
B1	Text-only transformer classifier	Single-signal reference
B2	Source-only credibility heuristic	Single-signal reference
B3	Equal-weight fusion of z1 to z5	Tests value of adaptive weights
B4	Retrieval-based fact-checking	Evidence-only reference
AEMRA	Proposed method	Full model

C. Metrics

We use Macro-F1 and AUROC for classification, expected calibration error (ECE) for score quality, NDCG@k for the ranking by Π, and latency per message. We report mean and standard deviation over [k] seeds, with a significance test.

D. Main results

Table V reports the main results on the test set.

TABLE V. MAIN RESULTS ON THE TEST SET

Method	MacroF1	AUROC	ECE	NDCG@10	Latency (ms)
B1 Text-only	[ ]	[ ]	[ ]	n/a	[ ]
B2 Source-only	[ ]	[ ]	[ ]	n/a	[ ]
B3 Equal-weight fusion	[ ]	[ ]	[ ]	[ ]	[ ]
B4 Retrieval factcheck	[ ]	[ ]	[ ]	n/a	[ ]
AEMRA	[ ]	[ ]	[ ]	[ ]	[ ]

E. Ablation and missing-signal tests

Table VI reports the ablation and missing-signal results.

TABLE VI. ABLATION AND MISSING-SIGNAL RESULTS

Setting	Macro-F1	AUROC	ECE
Full AEMRA	[ ]	[ ]	[ ]
Without AI-generation signal	[ ]	[ ]	[ ]
Without source credibility	[ ]	[ ]	[ ]
Without spatial consistency	[ ]	[ ]	[ ]
Without temporal consistency	[ ]	[ ]	[ ]
Without evidence support	[ ]	[ ]	[ ]
Fixed weights instead of gate	[ ]	[ ]	[ ]
30% of signals randomly missing	[ ]	[ ]	[ ]

XI. PERFORMANCE EVALUATION AND DISCUSSION

The results in Section X are analyzed along four dimensions. First, AEMRA is compared with the single-signal and equalweight baselines in macro-F1, AUROC, and ECE; AEMRA improves macro-F1 by [x] points over the strongest baseline (p = [value]). Second, the ablation shows that removing [signal] causes the largest drop ([x] macro-F1), and adaptive weights outperform fixed weights most on [disaster type or language]. Third, with 30% of signals missing, macro-F1 falls by [x] for AEMRA compared with [y] for equal-weight fusion, which supports the renormalization rule. Fourth, ranking by Π places [x]% of the most severe unreliable messages in the top ten (NDCG@10 = [value]) at a mean latency of [x] ms per message.

Explanation quality. Representative explanations, including one failure case, are shown in [figure or appendix]. A small moderator study measured [decision time and agreement] with and without explanations.

Threats to validity. Synthetic AI-generated messages may differ from real adversarial content. Detector accuracy shifts as generative models change. Authority registries and incident feeds differ by country, and retrieval depends on feed coverage. Severity, reach, and urgency scores are heuristics that need domain validation. Automated scoring should support, not replace, human moderators in life-safety decisions.

XII. CONCLUSION AND FUTURE SCOPE

This paper presented AEMRA, a reliability-aware framework that fuses AI-generation, source, spatial, temporal, and evidence signals with context-adaptive weights, tolerates missing signals, ranks messages by risk, and explains its verdicts. Experiments on [datasets] show that [summary of measured findings]. Future work includes multimodal messages with images and video, multilingual evaluation, online adaptation of thresholds, adversarial robustness against paraphrasing and evasion, and field trials with emergencymanagement agencies.

REFERENCES

[1] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters," in Proc. ICWSM, 2018.
[2] F. Alam et al., "HumAID: Human-annotated disaster incidents data from Twitter with deep learning benchmarks," in Proc. ICWSM, 2021.
[3] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018. [4] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on Twitter," in Proc. WWW, 2011. [5] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.
[6] E. Mitchell et al., "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in Proc. ICML, 2023.
[7] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, "FEVER: A large-scale dataset for fact extraction and verification," in Proc. NAACLHLT, 2018.
[8] A. Williams, N. Nangia, and S. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," in Proc. NAACLHLT, 2018.
[9] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020.
[10] N. Shazeer et al., "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in Proc. ICLR, 2017. [11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. ICML, 2017.

 ©IJSRED: All Rights are Reserved