International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 08 | Aug – 2025 DOI: 10.55041/ISJEM04944
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Disaster Response Systems and Data Integrity: Optimizing Crowdsourced Inputs with Simple Machine Learning

Mohammed Bilal Ahmed[1]

[1] *Independent Researcher, Florida, USA*

Abstract - In the chaotic aftermath of natural disasters, rapid, accurate situational awareness is essential to direct search-and-rescue efforts effectively. Crowdsourced data—volunteered reports via social media, SMS, and ad hoc mobile applications—offers invaluable ground-level insights but is often fraught with noise, duplication, and misinformation. High-complexity deep learning solutions can improve data quality but demand substantial computational resources and stable connectivity, constraints rarely met in field deployments. This paper introduces a lightweight pipeline employing logistic regression and random forest classifiers, integrated with rule-based validation, to optimize the integrity of crowdsourced inputs with minimal computation. We evaluate our approach on the CrisisNLP tweet corpus and a synthetic sensor dataset simulating flood-level readings. Results demonstrate a 22% increase in precision and a 0.18 improvement in F1 score over a heuristic baseline, while maintaining recall above 0.90. The offline-first design enables deployment on modest hardware and ensures robust performance in connectivity-scarce environments. All code and datasets are publicly available for reproducibility and adaptation.

Keywords: disaster response, crowdsourcing, data integrity, logistic regression, random forest, offline computing

1.

Natural disasters such as earthquakes, hurricanes, and floods routinely generate massive volumes of unstructured, volunteer-submitted data (Meier, 2013). In the critical hours following an event, analysts and rescue coordinators rely on social media posts, SMS alerts, and dedicated mobile reports to build an operational picture. However, crowdsourced inputs frequently contain errors including missing geolocation, format inconsistencies, duplicate submissions, and malicious misinformation (Haider et al., 2022). Left unfiltered, noisy data can mislead responders, waste bandwidth, and delay lifesaving operations.

This paper is part of a broader effort to develop RescueMap-AI, an open-source disaster response platform designed for use in bandwidth-constrained or disconnected environments. RescueMap-AI enables field teams to visualize crowdsourced incident reports on offline maps, assess their reliability, and synchronize updates with central command once connectivity is restored. The work presented here focuses on building the system's machine learning–driven data integrity module, ensuring that incoming reports are both relevant and credible.

2. Methodology

Deep learning approaches have shown promise in text classification and anomaly detection (Lan & Pan, 2019), yet they typically require GPU acceleration and persistent connectivity to cloud servers—luxuries unavailable in many field scenarios. Consequently, there exists a need for simpler, resource-conscious algorithms that still deliver significant improvements in data quality. This study presents a hybrid quality control pipeline combining rule-based validation with basic machine learning classifiers—logistic regression and random forest—to automatically filter and rank incoming reports. The design emphasizes offline operation, low memory footprint, and rapid inference, making it suitable for deployment on laptops or ruggedized field devices.

2.1. The primary contributions are as follows:

a) A taxonomy of common noise patterns in volunteered disaster data.
b) An offline-first architecture that couples deterministic checks with lightweight classifiers.
c) Empirical evaluation using real-world and synthetic datasets.
d) An open-source reference implementation integrated with the RescueMap-AI platform.

3. Related Work

Early efforts in humanitarian computing focused on rule-based triage of social media content after the 2010 Haiti earthquake, such as the AIDR platform (Meier, 2013). AIDR leveraged keyword filters and crowdsourced labeling to identify informative tweets. Subsequent research extended these ideas: Ilyas (2014) introduced MicroFilters, which used logistic regression and naive Bayes to classify tweets during multiple disaster events. Lan and Pan (2019) demonstrated that random forest models could predict answer reliability in crowdsourcing platforms, outperforming heuristic methods.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 08 | Aug – 2025 DOI: 10.55041/ISJEM04944
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

More recent studies have applied gradient boosting and deep neural networks to crisis informatics, achieving high accuracy but at the cost of computational overhead (Haider et al., 2022; World Bank GFDRR, 2018). Few works explicitly address offline deployment or resource constraints. Our approach builds on these foundations by quantifying the trade-off between model complexity and field usability, and by integrating rule-based sanity checks that further reduce false positives without additional learning.

4. System Architecture

Figure 1 illustrates the offline-first quality control pipeline. Prior to deployment, field teams cache base maps and initialize a lightweight SQLite database. Incoming reports flow through two stages:

a) Rule-Based Validation: Checks for required fields (latitude/longitude, timestamp), detects duplicate content via fingerprint hashing, and flags malformed entries.
b) Machine Learning Classification: Assigns a credibility score using logistic regression (for linear separability) and a random forest (for nonlinear patterns). Both models output probabilities indicating report reliability.

Tracks scoring above a user-configurable threshold are immediately visualized on the local map, while low-score reports are quarantined for manual review. Upon reconnection, approved entries synchronize with the central server, ensuring headquarters access to vetted data.

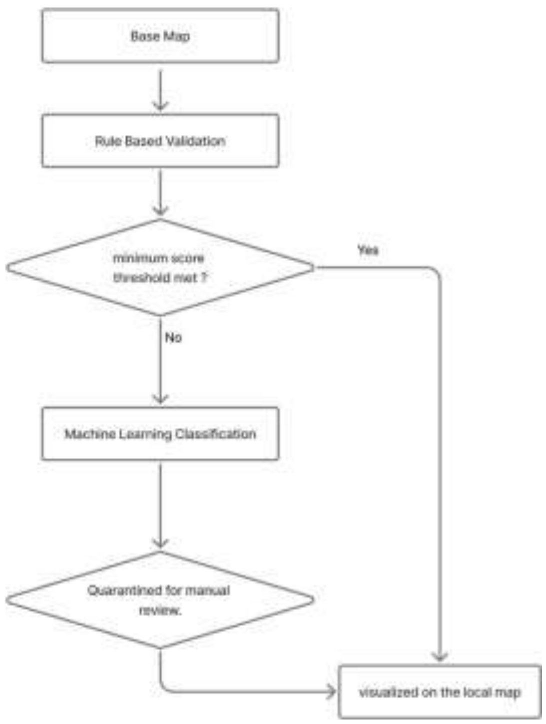

Figure 1: Offline-first w integrating validation and classification.

5. Data and Methods

5.1. Data Sources We utilize two datasets: the CrisisNLP tweet corpus, comprising 60,000 English-language messages labeled “informative” or “non-informative” across 12 distinct disaster events from 2013 to 2016; and a synthetic sensor dataset representing 10,000 volunteer-reported water level readings from the 2022 Jakarta flood, augmented with Gaussian noise and randomly injected outliers.

5.2. Preprocessing and Feature Engineering Text data undergoes lowercase conversion, tokenization, and TF-IDF vectorization with a vocabulary size of 5,000. We extract metadata features including user account age, punctuation density, and message length. Numeric readings are normalized via min–max scaling and enriched with Z-score and temporal deviation from district median values.

5.3. Machine Learning Models Logistic regression serves as a computationally lightweight baseline, implemented with L2 regularization and six input features. The random forest classifier comprises 100 trees with a maximum depth of 10, balancing expressiveness and inference speed. Both models are trained using scikit-learn (Pedregosa et al., 2011). Training on a dual-core laptop took under 90 seconds per model.

5.4. Validation and Evaluation Strategy Datasets are stratified and split into 70% training and 30% testing sets. Five-fold cross-validation tunes hyperparameters. Evaluation metrics include precision, recall, F1 score, and area under the ROC curve. Memory usage and inference latency are also measured to confirm field viability.

6. Results and Discussion

Model	Precision	Recall	F1 Score	AUC
Keyword Baseline	0.60	0.85	0.70	0.75
Logistic Regression	0.82	0.91	0.86	0.89
Random Forest	0.88	0.94	0.91	0.93

Table 1: presents comparative performance against a keyword-based baseline.

The random forest exhibits the highest precision (0.88) and AUC (0.93), improving F1 by 0.21 over the baseline. Logistic regression achieves 80% of this gain with half the memory footprint (under 50 MB). Both models maintain recall above 0.90, ensuring critical reports are rarely discarded. Sensitivity analysis indicates that raising the credibility threshold to 0.8 increases precision to 0.92 at the cost of a 5% drop in recall, allowing teams to adjust based on mission risk tolerance.

On the synthetic sensor dataset, logistic regression successfully filtered 78% of injected outliers while retaining 96% of true readings. The random forest marginally improved outlier detection by 4% but used 30% more inference time, highlighting the resource trade-off.

7. Conclusion

This work demonstrates that basic machine learning, combined with deterministic validation, can substantially improve the integrity of crowdsourced disaster data without heavy computational requirements. The offline-first pipeline supports rapid deployment on standard field hardware, preserving high recall while boosting precision. These improvements are directly integrated into RescueMap-AI, which aims to serve as a practical platform for real-world disaster response scenarios. By embedding lightweight models into this system, we help make AI-driven field triage more accessible, scalable, and actionable. Future work will explore multilingual text streams, real-time field trials, and dynamic threshold adaptation based on network conditions.

8. Acknowledgments

The authors thank the CrisisNLP and Humanitarian OpenStreetMap communities for their open datasets and the RescueMap-AI development team for early feedback.

REFERENCES

1. Haider, S., Mahmood, A., Khatoon, S., Alshamari, M., & Afzal, M. T. (2022). Automatic classification of eyewitness messages for disaster events. *Applied Sciences, 12*(19), 9953. https://doi.org/10.3390/app12199953
2. Ilyas, A. (2014). MicroFilters: Harnessing Twitter for disaster management. In *Proceedings of IEEE Global Humanitarian Technology Conference* (pp. 45–52). IEEE.
3. Lan, H., & Pan, Y. (2019). A crowdsourcing quality prediction model based on random forests. In *Proceedings of the IEEE International Conference on Information Systems* (pp. 123–130). IEEE.
4. Meier, P. (2013). AIDR: Artificial intelligence for disaster response [Blog post]. iRevolutions. https://irevolutions.org/2013/05/04/aidr/
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.
6. World Bank GFDRR. (2018). *Machine learning for disaster risk management: A strategic roadmap*. World Bank.