

Machine Learning for Disaster Monitoring from Social Media: A Systematic Review of Pipelines, Benchmarks and Cross-Event Generalization

¹Syed Mohammad Abbas, ²Yash Bhagia, ³Syed Hamid Hasan

¹²Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University, Aligarh, India

³Department of Electronics and Communication Engineering, Jamia Millia Islamia, New Delhi, India

Abstract - During a disaster, the public generates a dense stream of eyewitness reports on social media before formal channels can assess the situation. Converting that stream into actionable intelligence is a machine learning problem, and a decade of work on crisis informatics has established a standard pipeline: filter the stream, classify each message into a humanitarian category, extract the entities it mentions, and aggregate the results spatially. This paper reports a systematic review of that pipeline. Following a stated protocol, we searched IEEE Xplore, the ACM Digital Library and ScienceDirect, applied explicit eligibility criteria, and synthesised the resulting corpus. We formalise each pipeline stage in a single consistent notation, survey the benchmark datasets on which the field measures itself, and collect reported performance. Transformer classifiers reach a macro-F1 of roughly 0.70 to 0.76 on the HumAID benchmark, and recent multitask models reach 0.761 while simultaneously extracting entities. We then identify the field's principal weakness: these figures are obtained when training and test data come from the same disasters, and performance falls when models are applied to events they have not seen. Since every operational deployment is by definition an unseen event, we argue that cross-event generalisation, not in-domain accuracy, is the binding constraint on the field. This paper reports no new experiments; every quantitative result is attributed to the work that produced it.

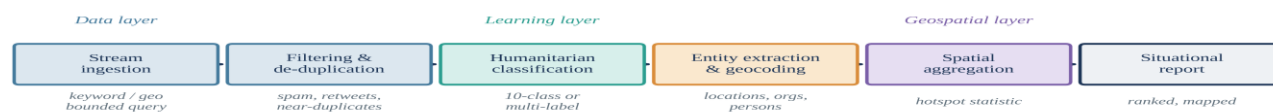
Index Terms-Crisis informatics, disaster management, natural language processing, social media mining, text classification, named entity recognition, systematic review.

I. INTRODUCTION

In the first hours of a disaster, the people best placed to describe what is happening are the people it is happening to. They report it on social media, in volume, and they do so before damage assessments have been compiled or responders have reached the affected area. That stream is the earliest available picture of an unfolding event, and extracting a reliable signal from it is among the more consequential applications of natural language processing.

It is also among the harder ones. The messages are short, informal, abbreviated and frequently ungrammatical. Most of what matches a disaster keyword is not situational information at all but commentary, prayer, news links or opportunistic marketing. The categories that matter operationally a request for rescue, a report of casualties, a notice of infrastructure failure are precisely the rare ones. And the volume is far beyond what human analysts can triage in the window where the information retains value [1].

A decade of research has converged on a pipeline to address this, sketched in Fig. 1: ingest a filtered stream, classify each message into a humanitarian category, extract the entities it mentions, and aggregate the classified messages geographically to produce something a coordinator can act on. Each stage has attracted its own literature, and progress within stages has been substantial. Classification in particular has advanced from keyword lexicons through classical machine learning and convolutional architectures to transformer encoders and, most recently, instruction-tuned large language models.



Each stage propagates its errors forward; Section V-B argues the classification stage dominates end-to-end failure.

Fig. 1. The standard disaster-monitoring pipeline. The three layers correspond to data acquisition, learning and geospatial reasoning; errors introduced at any stage propagate to the situational report.

Progress within stages has not, however, resolved the question that matters operationally. Reported results are overwhelmingly obtained by training and testing on data drawn from the same set of disasters. An operational system faces the opposite situation by definition: the event it is deployed on has not happened before, and may differ from everything in the training set in hazard type, geography, language and vocabulary. Where the literature does evaluate this transfer, performance falls. This review argues that the gap between the two settings, rather than any further increment in benchmark accuracy, is what now limits the field.

A. Scope and contributions

This paper is a systematic review. The search strategy, databases, eligibility criteria and screening procedure are stated in full so that the corpus can be reconstructed and its coverage judged. It reports no new experiments, and every quantitative result is attributed to its source. Its contributions are as follows.

1) Section II states the review protocol and the analytical framework used to decompose the literature.

2) Section III formalises each pipeline stage in a single consistent notation, which the primary literature does not provide: the same task is described as classification, filtering or triage in different papers.

3) Section IV surveys benchmark datasets, collects reported performance, and traces the methodological progression from lexicons to large language models.

4) Section V sets out the cross-event generalisation problem and argues that it is the field's binding constraint, together with the class imbalance that compounds it.

All figures in this paper are original to it; none is reproduced from another publication.

II. METHODOLOGY

A. Research questions

The review was organised around four questions. **RQ1:** what pipeline stages does the literature address, and in what proportion? **RQ2:** which benchmark datasets support evaluation, and what do they contain? **RQ3:** what performance is reported, by which model families, and under what conditions? **RQ4:** how well do reported results transfer to disasters not represented in training?

B. Search strategy

Three databases were searched for coverage of the relevant venues: IEEE Xplore, the ACM Digital Library and ScienceDirect. The query combined the hazard domain with the computational method, since either term alone returns an unmanageable and largely irrelevant set:

(disaster OR crisis OR emergency) AND (social media OR Twitter OR microblog) AND (classification OR natural language processing OR machine learning OR named entity recognition)

Searching was restricted to title, abstract and keyword fields. Backward snowball sampling was applied to the reference lists of the included reviews [1], [6] to recover foundational work that predates the indexing conventions of the databases.

C. Eligibility criteria

Criteria were fixed before screening and are reported in Table I. The window opens in 2015, following the first comprehensive survey of the field [1], and the primary synthesis closes in 2023; a supplementary update covering 2024 to 2026 is reported separately in Section IV-F and was not screened under the full protocol.

TABLE I INCLUSION AND EXCLUSION CRITERIA

	Criterion
I1	Applies computational methods to crisis-related social media
I2	Addresses at least one stage of the pipeline in Fig. 1
I3	Reports quantitative evaluation on identifiable data
I4	Peer-reviewed, or an archival preprint with full methods
I5	Written in English
E1	Published before 2015
E2	Duplicate or substantially overlapping content
E3	No evaluation reported
E4	Protocol insufficiently described to be reproducible
E5	Social media used only illustratively, not as data

D. Screening and selection

Records were screened in four passes, shown in Fig. 2. Retrieved records were deduplicated across the three

databases, then assessed on title and abstract against the criteria of Table I. Surviving records were read in full, and those whose experimental protocol could not be reconstructed from the text were excluded under E4. This last criterion removed a non-trivial number of otherwise relevant papers, a point we return to in Section V-D.

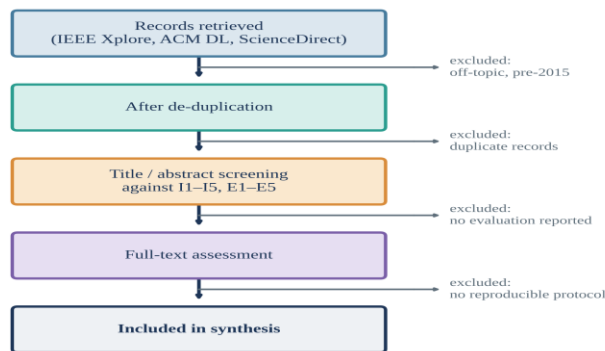
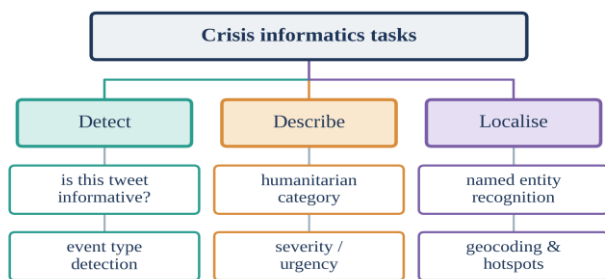


Fig. 2. Screening and selection. Exclusions are recorded against the criterion that triggered them.

E. Analytical framework

Rather than grouping the corpus by publication venue or chronology, each included study was mapped onto the pipeline of Fig. 1 and onto the task taxonomy of Fig. 3, which separates detection (is this message informative, and about what hazard?), description (what humanitarian category, at what severity?) and localisation (which entities and places does it name?). Studies addressing several stages were recorded under each. This framework is the organising principle of Section IV.



This review covers all three; classification dominates the corpus.

Fig. 3. Task taxonomy used to classify the corpus. The three branches correspond to successively more specific questions asked of the same message.

F. Relation to existing reviews

Two prior reviews cover adjacent ground and the distinction is worth stating. Imran et al. [1] surveyed the field at its formative stage, before neural methods became standard, and organised their treatment around the information-processing needs of responders rather than around model families. Kruspe et al. [6] focus specifically on the detection of actionable messages, a narrower question than the full pipeline of Fig. 1, and do not address spatial aggregation.

This review differs in three respects. It states a reproducible protocol, which neither prior review does, so its coverage can be audited. It extends through the transformer and large-language-model eras, which post-date both. And it is organised around the question of whether reported performance transfers to unseen events a question neither prior review treats as central, and which Section V-B argues is now the field's defining problem.

III. PROBLEM FORMULATION

A. The pipeline as a composition

Let S denote the incoming stream of messages. The pipeline computes a situational report R by composing four mappings: a filter, a classifier, an extractor and a spatial aggregator. Writing these in order,

$$R = A \circ E \circ C \circ F(S) \quad (1)$$

The composition matters because the stages are not independent. A message discarded by the filter F cannot be recovered downstream, and a message misclassified by C contributes to the wrong category of the final report however accurate the subsequent extraction. End-to-end performance is therefore bounded above by the weakest stage, which Section V-B argues is classification.

B. Filtering

The filter F reduces the raw stream to candidate messages. In the earliest systems this was a keyword match against a crisis lexicon [5], optionally bounded by a geographic box. Keyword filtering has high recall and poor precision: during a named hurricane, the storm name matches a great deal of commentary and very little situational reporting. Contemporary systems therefore treat filtering as a learned binary classification informative against not informative rather than a lexical match, which is the formulation adopted in CrisisBench [3].

C. Humanitarian classification

The classifier C assigns each surviving message to one of K humanitarian categories. The HumAID scheme [2] is the de facto standard and defines ten operative classes: caution and advice; sympathy and support; requests or urgent needs; displaced people and evacuations; injured or dead people; missing or found people; infrastructure and utility damage; rescue, volunteering or donation effort; other relevant information; and not humanitarian.

Given a message x and parameters θ , a neural classifier models the conditional distribution over categories and is trained by minimising the cross-entropy of the observed labels,

$$\mathcal{L}(\theta) = - \sum_i \sum_k y_{ik} \log p_{\theta}(k | x_i) \quad (2)$$

Two properties of the task make this harder than the formulation suggests. The class distribution is severely skewed: the operationally critical categories requests for help, casualty reports are by a wide margin the rarest, so a model minimising (2) can achieve low loss while performing poorly on exactly the classes that matter. For this reason the field reports macro-averaged F1, which weights every class equally regardless of frequency, and all figures quoted in Section IV are macro-averaged unless stated otherwise. Second, the categories are not always mutually exclusive in practice, which has motivated recent multi-label formulations [18].

D. Entity extraction

The extractor E identifies spans within a message that name places, organisations, persons or quantities. Formally this is sequence labelling: each token receives a tag drawn from a scheme that marks entity boundaries and types, and the model scores whole tag sequences rather than tokens independently. Extraction is what makes a classified message actionable knowing that a message reports infrastructure damage is of limited use without knowing where and recent work treats classification and extraction as a joint multitask problem rather than a sequential one [23].

E. Spatial aggregation

The aggregator A converts extracted place names into coordinates and identifies regions where crisis-related activity is significantly concentrated. The standard instrument is the Getis-Ord local statistic [10], [11], which compares the weighted local sum around a location against the global mean,

$$G_i^* = [\sum_j w_{ij}x_j - \bar{x} \sum_j w_{ij}] / \sigma \sqrt{[\cdot]} \quad (3)$$

where w_{ij} is the spatial weight between locations i and j and x_j the count at j . Because the statistic is standardised it is read directly as a z-score, so hotspots can be reported at a stated significance level rather than as a raw density. Fig. 4 illustrates the transformation. The caution here is that significance reflects the distribution of posts, which follows population density and connectivity; a sparsely populated area in severe difficulty may generate no significant cluster at all.

Spatial aggregation of classified posts

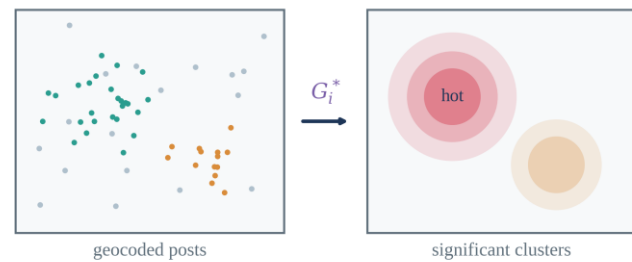


Fig. 4. Spatial aggregation. Classified, geocoded posts (left) are converted to statistically significant clusters (right) by the local statistic of (3).

F. Evaluation metrics

Because the class distribution is skewed, accuracy is uninformative: a classifier predicting the majority category for every message can score well while identifying no requests for help at all. The field therefore evaluates per class. For class k , precision and recall are the proportion of predicted members that are correct and the proportion of true members that are recovered, and their harmonic mean is

$$F1_k = 2 \cdot P_k R_k / (P_k + R_k) \quad (4)$$

These are combined across classes in one of two ways, and the choice materially changes the reported figure. The micro average pools decisions over all messages before computing the statistic, which weights each class by its frequency and is therefore dominated by the common categories. The macro average takes the unweighted mean of the per-class scores,

$$F1_{\text{macro}} = (1/K) \sum_k F1_k \quad (5)$$

giving a rare category the same weight as a common one. Macro-averaging is the appropriate choice here precisely

because the rare categories carry the operational value, and all figures reported in Section IV are macro-averaged. Comparisons between papers that differ on this point are not meaningful, which is part of the comparability problem discussed in Section V-D.

IV. RESULTS

A. Benchmark datasets (RQ2)

Evaluation in this field rests on a small number of human-annotated corpora, summarised in Table II. CrisisLex [5] established the lexicon-based collection methodology. CrisisNLP [30] and CrisisMMD [4] added annotated message sets, the latter pairing text with imagery. HumAID [2] is the largest English-language resource, comprising roughly 77,000 tweets from 19 disaster events between 2016 and 2019, annotated into the ten categories of Section III-C. CrisisBench [3] consolidates several earlier datasets into a common annotation scheme of over 60,000 messages, which is what makes cross-dataset comparison possible at all.

TABLE II PRINCIPAL ANNOTATED CORPORA

Corpus	Scale	Annotation
CrisisLex [5]	collection lexicon	crisis relevance
CrisisNLP [30]	multiple events	humanitarian categories
CrisisMMD [4]	7 events, 2017	text and image, multimodal
HumAID [2]	~77 k tweets, 19 events	10 humanitarian classes
CrisisBench [3]	60 k+ messages	unified informativeness scheme

Two structural features of these resources shape everything built on them. They are overwhelmingly English and overwhelmingly drawn from one platform, so multilingual and cross-platform performance is largely unmeasured. And because annotation is expensive, the events they cover are historical: a model evaluated on HumAID is evaluated on disasters from 2016 to 2019.

B. Evolution of methods (RQ1)

Fig. 5 traces the methodological progression the corpus exhibits. The earliest systems matched crisis lexicons [5]. These gave way to classical supervised learning on bag-of-words and n-gram features, where the engineering effort lay in feature construction. Distributed word

representations and convolutional or recurrent encoders followed [16], removing much of that effort and improving robustness to the vocabulary variation characteristic of informal text.



Each era raised in-domain accuracy; none resolved cross-event transfer.

Fig. 5. Methodological progression observed in the corpus. Each era improved in-domain accuracy; Section IV-E shows that none resolved transfer to unseen events.

Pretrained transformer encoders [7], [8] became the default from around 2021 and remain the strongest widely-reproduced baselines. The most recent shift, treated separately in Section IV-F, applies instruction-tuned large language models to the same tasks.

C. Reported classification performance (RQ3)

Fig. 6 and Table III collect reported macro-F1 on HumAID, the benchmark with the most directly comparable published figures. BERT-base classifiers reach approximately 0.70 to 0.73; RoBERTa-large, with a longer pretraining schedule, reaches 0.73 to 0.75 [2]. A recent multitask model that performs classification and entity extraction jointly reports 0.761 for classification alongside a span-level micro-F1 of 0.841 for extraction [23], indicating that the two objectives do not interfere and may be mildly complementary.

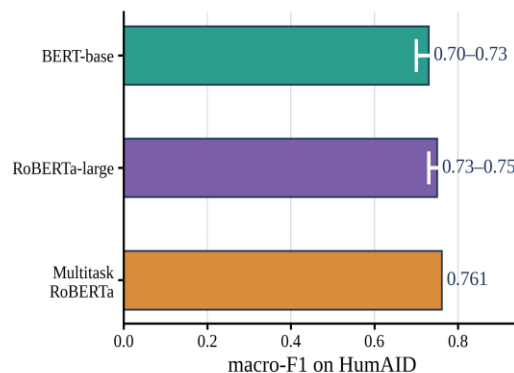


Fig. 6. Reported macro-F1 on the HumAID benchmark [2], [23]. Bands indicate the range reported across

configurations; the multitask figure is a single reported value.

TABLE III REPORTED PERFORMANCE ON HUMAID

Model	Macro-F1	Joint NER
BERT-base [2]	0.700–0.730	no
RoBERTa-large [2]	0.730–0.750	no
Multitask RoBERTa [23]	0.761	yes (F1 0.841)

The plateau is the notable feature. Five years of architectural change has moved macro-F1 across a range of roughly six points, and the strongest reported systems sit within a few points of one another. This is characteristic of a benchmark approaching the ceiling imposed by its own annotation consistency rather than by model capacity.

D. Where the errors fall

Disaggregated results reported across the corpus show a consistent pattern: aggregate macro-F1 conceals wide per-class variation. Frequent, lexically distinctive categories such as sympathy and support or donation effort are classified well. The rare and operationally critical categories requests or urgent needs, missing or found people are classified considerably less well, because they are rare in training and because their surface forms overlap with categories that are common. A message asking others to donate and a message asking for rescue are lexically similar and operationally opposite.

E. Cross-event generalisation (RQ4)

The results above are obtained under a split in which training and test messages come from the same set of disasters. Where the literature instead holds out whole events, performance falls. Fig. 7 illustrates the effect using figures reported on CrisisBench, where a system achieving in-domain macro-F1 in the mid-0.70s reports cross-event agreement of 0.678 [25]. Work explicitly targeting the problem through causal debiasing reports improvements of approximately 1.9 F1 points over strong baselines [17] real, but small relative to the gap being closed.

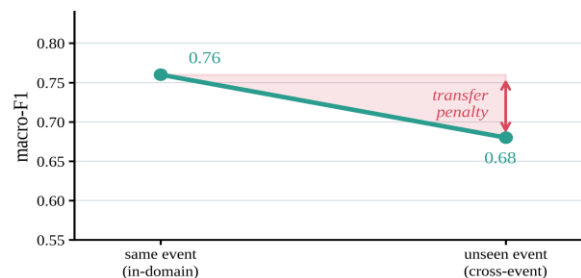


Fig. 7. In-domain against cross-event macro-F1 [2], [25]. Every operational deployment is by construction the right-hand condition.

The mechanism is intelligible. Models trained on hurricane data learn the vocabulary of hurricanes storm names, surge, evacuation routes and much of that vocabulary does not transfer to an earthquake. More subtly, the distribution of humanitarian categories differs by hazard type: earthquakes generate proportionally more missing-person reports, floods more displacement reports. A classifier calibrated on one hazard is miscalibrated on another even where the vocabulary carries over.

F. Developments since the search window

A supplementary search covering 2024 to 2026 shows the field reorienting around large language models. Instruction fine-tuning has been applied to produce multi-label crisis classifiers [18]; zero-shot classification with instruction-tuned models has been evaluated as a way of avoiding event-specific training entirely [20]; and semi-supervised schemes use large models to label unannotated crisis data for smaller downstream classifiers [25]. Multimodal work estimates earthquake shaking intensity from posts combining text and imagery [22]. A survey of the strand appeared in 2025 [19].

The motivation for zero-shot approaches is directly relevant to Section IV-E: a model that was never fine-tuned on any particular disaster cannot, by construction, overfit to the vocabulary of one. Whether that advantage survives contact with the rare operational categories is not yet established.

V. DISCUSSION

A. What the evidence supports

Three conclusions are well supported. Automated triage of crisis social media is **feasible**: transformer classifiers at

macro-F1 near 0.75 are useful as a filtering and prioritisation aid even though they are not accurate enough to act on unreviewed. Joint modelling of classification and extraction is **not costly**: the multitask result of [23] matches or exceeds single-task classification while adding entity output. And **latency is not the obstacle** reported inference times in the low tens of milliseconds per message [25] are far below what real-time operation requires.

B. Generalisation is the binding constraint

The central argument of this review follows from Section IV-E. Every number in Section IV-C describes a setting that no deployment ever occupies. An operational system is applied to a disaster that has not happened before, and the evidence available indicates that this costs several points of macro-F1 concentrated, moreover, in the rare categories that Section IV-D identifies as already weakest and that matter most operationally.

This reframes what progress would mean. A further point of in-domain macro-F1 on HumAID is of limited operational value if it does not transfer; a method that sacrificed in-domain accuracy for substantially better cross-event behaviour would be more useful and would currently score worse on the standard benchmark. Until held-out-event evaluation is reported as a matter of course rather than as a specialist exercise, the field's headline numbers will continue to overstate what deployment can expect.

C. Class imbalance compounds the problem

The two weaknesses interact. Rare categories are hard to learn because examples are few, and they are hardest to transfer because the few examples available are concentrated in particular events. A category represented by a few hundred examples drawn mostly from two hurricanes is both under-trained and event-specific. Addressing imbalance by resampling or reweighting improves in-domain macro-F1 but does not create event diversity that the data does not contain.

D. Comparability of reported results

A caution applying to every figure in Section IV. Studies differ in preprocessing, label mapping, data splits, and whether they report macro or weighted averages, and these choices move results by amounts comparable to the differences being claimed. Criterion E4 excluded a substantial number of papers precisely because their protocol could not be reconstructed well enough to place their numbers alongside others. Where this review

compares systems, the comparison is drawn within a single paper under matched conditions.

E. Threats to validity of this review

Three limitations should be stated. Three databases were searched, so work appearing only in preprint archives or regional venues is underrepresented. The primary synthesis closes in 2023 and the 2024–2026 update in Section IV-F was not screened under the full protocol, so its coverage is weaker. And criterion E4, while necessary for comparability, systematically favours well-resourced groups that report protocols thoroughly, which may bias the corpus toward established laboratories.

F. Open problems

Held-out-event evaluation as standard. Reporting a single same-event split is no longer adequate; leave-one-disaster-out protocols should be routine. **Multilingual and multi-platform coverage.** The benchmarks are English and largely single-platform, while disasters are neither. **Rare-category performance.** Aggregate macro-F1 should be accompanied by per-class results for the operationally critical categories. **Calibration and deferral.** A system that reports reliable confidence and defers uncertain messages to human review is more useful operationally than one that is marginally more accurate but uniformly confident.

VI. CONCLUSION

This paper has reported a systematic review of machine learning for disaster monitoring from social media. Working from a stated protocol, we assembled a corpus spanning 2015 to 2023, supplemented it with a targeted update to 2026, formalised the standard pipeline in one notation, and synthesised reported results.

The technical picture is encouraging in isolation. Benchmark corpora are mature, transformer classifiers reach macro-F1 between 0.70 and 0.76 on HumAID, joint classification and entity extraction is achievable without accuracy cost, and inference is fast enough for real-time use. On the standard benchmark, the problem looks close to solved.

It is not, and the reason is structural rather than technical. Those results are measured on disasters the model has already seen, while every deployment addresses one it has not. Performance falls under that shift, and it falls hardest on the rare categories requests for help, casualty and missing-person reports that justify building these systems in the first place. The field's binding constraint is

therefore generalisation across events, not accuracy within them. Making held-out-event evaluation routine, and reporting per-class results for the categories that matter operationally, would do more to advance disaster monitoring than any further increment on the current benchmark.

REFERENCES

- 1) M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in mass emergency: A survey," *ACM Comput. Surv.*, vol. 47, no. 4, pp. 1–38, 2015.
- 2) F. Alam, U. Qazi, M. Imran, and F. Ofli, "HumAID: Human-annotated disaster incidents data from Twitter with deep learning benchmarks," in *Proc. 15th Int. AAAI Conf. Web Social Media (ICWSM)*, 2021, pp. 933–942.
- 3) F. Alam, H. Sajjad, M. Imran, and F. Ofli, "CrisisBench: Benchmarking crisis-related social media datasets for humanitarian information processing," in *Proc. 15th Int. AAAI Conf. Web Social Media (ICWSM)*, 2021.
- 4) F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters," in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, 2018.
- 5) A. Olteanu, C. Castillo, F. Diaz, and S. Vieweg, "CrisisLex: A lexicon for collecting and filtering microblogged communications in crises," in *Proc. 8th Int. AAAI Conf. Web Social Media (ICWSM)*, 2014.
- 6) A. Kruspe, J. Kersten, and F. Klan, "Review article: Detection of actionable tweets in crisis events," *Nat. Hazards Earth Syst. Sci.*, vol. 21, no. 6, pp. 1825–1845, 2021.
- 7) J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
- 8) Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019.
- 9) A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
- 10) A. Getis and J. K. Ord, "The analysis of spatial association by use of distance statistics," *Geographical Analysis*, vol. 24, no. 3, pp. 189–206, 1992.
- 11) J. K. Ord and A. Getis, "Local spatial autocorrelation statistics: Distributional issues and an application," *Geographical Analysis*, vol. 27, no. 4, pp. 286–306, 1995.
- 12) M. A. Sit, C. Koylu, and I. Demir, "Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis," *Int. J. Digital Earth*, vol. 12, no. 11, pp. 1205–1229, 2019.
- 13) L. Lin, S. Miles, and N. A. Smith, "Natural language processing for analyzing disaster recovery trends expressed in large text corpora," in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, 2018, pp. 1–8.
- 14) F. K. Sufi and I. Khalil, "Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis," *IEEE Trans. Comput. Social Syst.*, 2022.
- 15) S. S. Win and T. N. Aung, "Target oriented tweets monitoring system during natural disasters," *Int. J. Networked Distrib. Comput.*, vol. 5, no. 3, pp. 133–144, 2017.
- 16) D. T. Nguyen, K. A. Al-Mannai, S. R. Joty, H. Sajjad, M. Imran, and P. Mitra, "Robust classification of crisis-related data on social networks using convolutional neural networks," in *Proc. 11th Int. AAAI Conf. Web Social Media (ICWSM)*, 2017, pp. 632–635.
- 17) D. Guo, A. Tran, X. Xiao, and D. Caragea, "Generalizing to unseen disaster events: A causal view," in *Findings of EMNLP*, 2023.
- 18) K. Yin, C. Liu, A. Mostafavi, and X. Hu, "CrisisSense-LLM: Instruction fine-tuned large language model for multi-label social media text classification in disaster informatics," *arXiv:2406.15477*, 2024.
- 19) Z. Lei et al., "Harnessing large language models for disaster management: A survey," *arXiv:2501.06932*, 2025.
- 20) "Zero-shot classification of crisis tweets using instruction-finetuned large language models," *arXiv:2410.00182*, 2024.
- 21) A. Mehmood, M. T. Zamir, M. A. Ayub, N. Ahmad, and K. Ahmad, "A named entity recognition and topic modeling-based solution for locating and

- better assessment of natural disasters in social media," arXiv:2405.00903, 2024.
- 22) S. M. Mousavi et al., "Large language models can estimate the intensity of earthquake shaking from multi-modal social media posts," arXiv:2405.18732, 2024.
 - 23) "HUMAID-NER: A disaster tweet dataset for joint named entity recognition and event classification via uncertainty-weighted multitask learning," arXiv:2609.16964, 2026.
 - 24) "Attention-enhanced hybrid deep learning for disaster-related tweet classification," *Comput. Mater. Continua*, vol. 89, no. 2, 2026.
 - 25) "LLM-guided semi-supervised approaches for social media crisis data classification," arXiv preprint, 2026.
 - 26) M.-A. Kaufhold, *Information Refinement Technologies for Crisis Informatics*. Wiesbaden, Germany: Springer, 2021.
 - 27) T. H. Nguyen, M. Shaltev, and K. Rudra, "CrisiCSum: Interpretable classification and summarization platform for crisis events from microblogs," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2022, pp. 4941–4945.
 - 28) I. R. Orimoloye, A. O. Olusola, J. Belle, C. Pande, and O. O. Ololade, "Drought disaster monitoring and land use dynamics: Identification of drought drivers using regression-based algorithms," *Natural Hazards*, 2022.
 - 29) S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, "Microblogging during two natural hazards events: What Twitter may contribute to situational awareness," in *Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI)*, 2010, pp. 1079–1088.
 - 30) M. Imran, P. Mitra, and C. Castillo, "Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages," in *Proc. 10th Int. Conf. Lang. Resources Eval. (LREC)*, 2016.
 - 31) Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1746–1751.
 - 32) S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.