THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 6. Issue 1 (2026)
https://doi.org/10.62019/zabvxd97

ASIAN BULLETIN OF BIG DATA MANAGEMENT
http://abbdm.com/
ISSN (Print): 2959-0795
ISSN (online): 2959-0809

Humaid-Ner: A Disaster Tweet Dataset for Joint Named Entity Recognition and Event Classification Via Uncertainty-Weighted Multitask Learning

Aijaz Ali*, Nazish Basir, Sarfaraz Nawaz, Danish Nazir Arain, Haris Ali

Chronicle

Article history
Received: Jan 2, 2026
Received in the revised format: Jan 13, 2026
Accepted: Feb 17 2026
Available online: March 9, 2026

Aijaz Ali*, & Sarfaraz Nawaz are currently affiliated with Department of Software Engineering, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan.
Email: aijaz.laghari@students.usindh.edu.pk
Email: sarfaraz.mangi@students.usindh.edu.pk

Nazish Basir Is currently affiliated with Department of Information Technology, Faculty of Engineering and Technology, University of Sindh, Jamshoro, Pakistan.
Email: nazish.basir@usindh.edu.pk

Danish Nazir Arain is currently affiliated with Dr. A. H. S. Bukhari Postgraduate Centre Of ICT, University of Sindh, Jamshoro, Pakistan, Pakistan
Email: danish.arain@usindh.edu.pk

Haris Ali is currently affiliated with Department of Software Engineering,
Mehran University of Engineering & Technology,
Jamshoro, Pakistan.
Email: harislag77@gmail.com

Corresponding Author*

Abstract

Rapid extraction of structured information from social media is central to effective humanitarian response, yet disaster tweet resources to date offer only document-level category labels with no span-level entity annotations. We address this gap with HUMAID-NER, the first named entity recognition dataset built on the HumAID benchmark: 60,000 English disaster tweets annotated in BIO format across ten operationally motivated entity types, including CASUALTY, DISPLACED, REQUEST, RESOURCE, and RESCUE, yielding 21 entity classes and roughly 175,000 labelled entity spans. Annotations were produced through a reproducible three-stage hybrid pipeline that combines a spaCy transformer backbone, disaster-domain EntityRuler patterns, and structured regular expressions with priority-based overlap resolution. We also propose a joint multitask learning framework that performs disaster-specific NER and humanitarian event classification through a single RoBERTa-large encoder. A core difficulty in joint training is task-conflict: the NER objective produces up to 2,688 token-level gradient signals per example while classification contributes one, and under fixed task weights this imbalance caused classification macro-F1 to fall 1.4 points across epochs. Homoscedastic uncertainty weighting with learnable per-task log-variance parameters resolves the conflict, paired with a two-stage training schedule that freezes the lower 18 of 24 encoder layers in the second stage to permit task-specific specialisation without eroding shared representations. A controlled four-row ablation study isolates each component's contribution. On the HUMAID-NER validation set, the proposed system reaches NER span micro-F1 of 0.841 and classification macro-F1 of 0.761 simultaneously; under this setting, classification performance meets or exceeds dedicated single-task RoBERTa-large classifiers on the same benchmark (0.730–0.750), suggesting joint modelling introduces no classification trade-off while adding complete entity extraction capability. A real-time web dashboard demonstrates end-to-end deployment. Dataset, models, and pipeline code are released to support reproducibility and future crisis informatics research.

Keywords: Disaster Tweet Analysis, Named Entity Recognition, Multitask Learning, Uncertainty Weigh, RoBERTa, Crisis Informatics.

INTRODUCTION

When a major disaster hit, affected communities rely on social media almost immediately. Evacuation requests, casualty reports, and resource needs appear within minutes of an event, generating a real-time information stream that no structured reporting system can replicate (Imran et al., 2015). For humanitarian organisations this create a critical operational problem: the sheer volume and noise make manual monitoring impossible at crisis pace, yet buried within that stream is exactly the actionable content that response coordinators need. NLP systems that can filter, classify, and pull structured information out of this stream are not an academic exercise; they are a practical requirement for modern disaster response (Ocal & Torun, 2025). Progress in classifying disaster-related social media by humanitarian category has been considerable. The HumAID dataset, (Alam et al., 2021) in 2021, provides roughly 77,000 human-labelled English tweets from 19 major natural disaster events across ten humanitarian classes, and transformer-based classifiers on this benchmark have set strong performance standards. The CrisisNLP corpora (Imran et al., 2016) similarly showed, at an earlier stage, that informational tweet categories can be reliably identified through supervised deep learning. Classification alone, however, does not fully satisfy operational requirements. Knowing that a tweet belongs to the *injured or dead people* category tells a coordinator what type of message it is, it does not say *where* casualties occurred, *how many* are reported, or *what specific resources* are being sought. Those answers come from named entity recognition (NER): direct extraction of typed spans from the tweet text using transformer-based models (Akpan, 2024; Munyao & Ndia, 2025).

No existing dataset, to our knowledge, provides NER annotations designed specifically for the disaster domain. Nor has any prior work trained a single model to simultaneously handle disaster-specific NER and disaster event classification. Social media NER resources inherit generic entity taxonomies from newswire benchmarks, omitting operationally critical types such as CASUALTY, DISPLACED, RESOURCE, RESCUE, and REQUEST. Running separate models for each task doubles inference cost and forfeits the representational synergy that shared disaster vocabulary naturally provides (Rijhwani et al., 2020; Ritter et al., 2011). The gap is missing annotated data plus missing joint modelling capability; is the central problem this work takes on.

Joint multitasks learning (MTL) with a shared encoder is, in principle, well-suited to this setting (X. Liu et al., 2019). NER and classification, however, differ fundamentally in gradient structure. NER generate dense token-level supervision: 21 entity classes per token, producing up to 2,688 signals per example. Classification produces one sentence-level signal per example. Under fixed-weight loss combinations, the heavier NER gradients systematically dominate optimisation, causing the classification head to underfit beyond the early epochs. In our controlled experiments; classification macro-F1 peaked at epoch three and dropped by 1.4 percentage points by epoch fifteen, a progressive, structural failure rather than noise. This instability rules out fixed-weight joint training for reliable deployment.

Our solution pairs homoscedastic uncertainty-based loss weighting (Kendall et al., 2018; Zhang et al., 2025) which replaces static task weights with learnable log-variance parameters, with a two-stage training procedure adapted from MT-DNN. Stage one builds shared cross-task representations across all layers; stage two freezes the bottom eighteen encoder layers and lets each task head specialise using the remaining capacity. The backbone is RoBERTa-large (Conneau et al., 2020; Y. Liu et al., 2019), chosen for its stronger pretraining recipe and confirmed compatibility with TPU v3-8. To support all of this, we extend HumAID with disaster-specific NER annotations across ten BIO-format entity types, producing HUMAID-NER, the first NER-annotated extension of this widely-cited benchmark.

The main contributions of this work are:

- We introduce **HUMAID-NER**, 60,000 tweets extending HumAID with BIO-format NER annotations across ten disaster-specific entity types, balanced across ten humanitarian classification labels, constituting the first NER-annotated resource built on the HumAID benchmark.

- We propose a **joint multitask framework** combining RoBERTa-large, Kendall uncertainty weighting , and two-stage layer-freezing training . A controlled four-row ablation isolates each component's contribution and shows that uncertainty weighting prevents the classification degradation that fixed-weight training produces.
- We deploy the joint model as a **real-time web dashboard** for disaster response support, providing simultaneous entity extraction and event classification from live tweet input and demonstrating end-to-end applicability beyond academic evaluation.

Section II reviews related work. Section III describes the dataset and model. Section IV presents experimental results, component analysis, and the deployment dashboard. Section V concludes.

LITERATURE REVIEW

Using social media as a real-time situational awareness source during crises has been an active research area since the early 2010s. A survey (Imran et al., 2015) NLP methods for crisis messaging and identified humanitarian information classification as the field's central computational challenge. HumAID (Alam et al., 2021)stands as the most comprehensive English-language disaster tweet resource available today, covering 19 disasters (2016–2019) with roughly 77,000 tweets labelled across ten humanitarian categories; transformer-based systems on this benchmark achieved macro-F1 of 0.70–0.75, establishing the classification baseline this work builds on directly. Alam et al. also extended the disaster tweet line with CrisisMMD (Alam et al., 2018), a multimodal dataset pairing tweet text with images from seven 2017 events. That work, like HumAID and CrisisNLP , is limited to document-level categorical labels with no span-level entity annotation. Broader work on disaster tweet categorisation for operational use (Stowe et al., 2016) and automated geo-event mapping from social streams (Fan et al., 2020) further underlines the need for span-level extraction alongside event classification.

CoNLL-2003 (Sang & De Meulder, 2003) formalised NER evaluation around four entity types suited to newswire text, and those categories have dominated benchmarks ever since. The work by (Ritter et al., 2011) documented just how poorly standard NLP pipelines transfer to tweets: POS tagging accuracy fell from 0.97 to 0.80, making tweet-specific models a practical requirement. The gap narrowed considerably with transformer fine-tuning , which reduced dependence on large task-specific corpora through pretrained contextual representations, superseding the contextual embedding approaches of (Prabha & Sardana, 2025). TweetNER7 (Ushio et al., 2022), a dedicated Twitter NER benchmark with seven entity types across 11,382 English tweets, reflects the community's continued investment in this problem. In another study (Rijhwani et al., 2020) extended neural NER to low-resource settings via soft gazetteers, incorporating cross-lingual entity knowledge.

HUMAID-NER draws on this principle: all ten entity types are grounded in the operational vocabulary of humanitarian response rather than inherited from general-purpose newswire categories. Joint training on related tasks improves generalization (Feng et al., 2022), and (S. Chen et al., 2024) later surveyed the extensive NLP literature that followed. For disaster tweets in particular, NER and event classification share heavy vocabulary overlap, *collapsed, shelter, trapped, evacuation* appear prominently in both task distributions, which makes joint training well-motivated on theoretical and empirical grounds. MT-DNN (X. Liu et al., 2019) showed this concretely: a single BERT encoder jointly trained on sentence-level and token-level NLP tasks consistently outperforms single-task fine-tuning, with lower encoder layers learning universal linguistic features while upper layers encode task-specific patterns. That layer-function insight directly motivates the two-stage training procedure used here. Combining individual task losses into a single training objective is a recurring challenge in MTL. (S. Chen et al., 2024) identified gradient imbalance as a primary driver of negative transfer: one task's gradients simply overpower shared parameter updates. A study by (Kendall et al., 2018) addressed this with homoscedastic uncertainty weighting, giving each task a learnable log-variance parameter that scales its loss contribution dynamically throughout training.

The log-variance parameterisation prevents collapse to trivial solutions where $\sigma \rightarrow \infty$, and the approach transfers cleanly from its original computer vision setting to the combination of token-level NER and sentence-level classification studied here. PCGrad (Xin et al., 2022) offers an alternative that projects conflicting gradients to eliminate destructive interference, but the approach roughly doubles memory cost, prohibitive at our training scale. GradNorm (Z. Chen et al., 2018) provides a third option via dynamic gradient magnitude scaling, though it similarly increases per-step compute. Uncertainty weighting introduces only two scalar parameters with negligible overhead, making it the practical choice for our setup. All three methods are complementary, and a direct comparison on disaster-domain MTL remains an open direction.

METHODOLOGY

A. Dataset Description

HumAID provides 77,637 English tweets spanning 19 major disasters (2016–2019), each labelled with one of ten humanitarian categories but carrying no span-level entity annotations. We extended a balanced 60,000-tweet subset with full BIO named entity annotations across ten disaster-specific entity types to produce HUMAID-NER. The dataset is partitioned into training (72%), validation (14%), and test (14%) splits, stratified by humanitarian category to preserve label distribution. Table 1 reports the statistics.

Table 1.
HUMAID-NER Dataset Statistics

Split	Tweets	Entity Spans	Avg/Tweet
Train (72%)	43,200	~126,000	2.91
Validation (14%)	8,400	~24,500	2.93
Test (14%)	8,400	~24,500	2.92
Total	**60,000**	**~175,000**	**2.92**

The entity taxonomy defines ten types rooted in the operational vocabulary of humanitarian response rather than general-purpose newswire categories : LOCATION, CASUALTY, DISPLACED, REQUEST, RESOURCE, RESCUE, DISASTER_TYPE, ORGANIZATION, PERSON, and NUMBER, yielding 21 BIO classes (one O class and B-/I- prefixes for each type). Table 2 gives definitions and examples for each type.

Manually annotating 60,000 tweets was not feasible, so we built a three-stage hybrid auto-labelling pipeline, illustrated in Figure. 1. Stage 1 runs the spaCy (Honnibal et al., 2020) transformer model en_core_web_trf to produce base entity predictions for PERSON, LOCATION, ORGANIZATION, and DATE. Stage 2 passes the text through an

EntityRuler component loaded with domain-specific rules covering named disasters, humanitarian organisations, and key disaster-vocabulary phrases. Stage 3 applies regular expressions to capture the remaining structured types: CASUALTY, DISPLACED, REQUEST, RESOURCE, and RESCUE.

Table 2.
HUMAID-NER Entity Taxonomy

Type	Description	Example
LOCATION	Geographical references	“Puerto Rico”
DISASTER_TYPE	Named hazard or event type	“Hurricane Maria”
CASUALTY	Injury/fatality expressions	“17 dead”
DISPLACED	Evacuation/displacement counts	“1,000 evacuees”
REQUEST	Expressed need for aid	“need food”
RESOURCE	Available aid supplies	“water trucks”
RESCUE	Active emergency operations	“rescue teams”
ORGANIZATION	Named responding organisations	“Red Cross”
PERSON	Named individuals in tweets	“Mayor Cruz”
NUMBER	Standalone numerical values	“Day 3”

Where two spans compete, the longer one is retained; for equal-length conflicts, neural predictions take priority over regex matches. All annotations are then converted to BIO format via offset mapping from the RoBERTa tokeniser, and continuation subword tokens receive label -100 so they are excluded from loss computation.

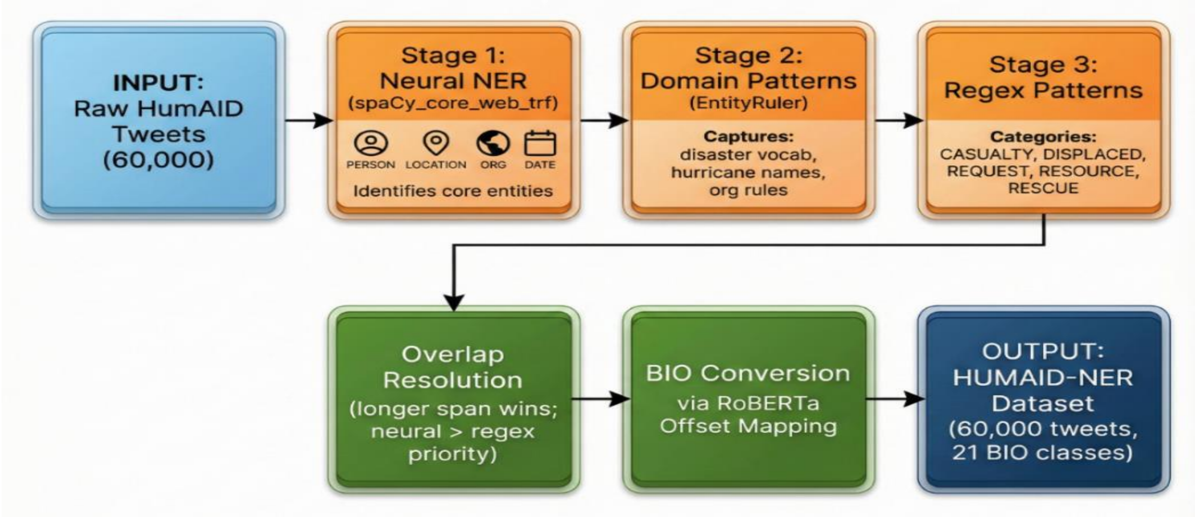

INPUT: Raw HumAID Tweets (60,000)
Stage 1: Neural NER (spaCy_core_web_trf)
PERSON
LOCATION
ORG
DATE
Identifies core entities
Stage 2: Domain Patterns (EntityRuler)
Captures: disaster vocab, hurricane names, org rules
Stage 3: Regex Patterns
Categories: CASUALTY, DISPLACED, REQUEST, RESOURCE, RESCUE
Overlap Resolution (longer span wins; neural > regex priority)
BIO Conversion via RoBERTa Offset Mapping
OUTPUT: HUMAID-NER Dataset (60,000 tweets, 21 BIO classes)

Figure 1.
HUMAID-NER construction pipeline. Stage 1: spaCy transformer for base NER. Stage 2: disaster-domain EntityRuler patterns. Stage 3: regex for structured entity types. Overlap resolution and BIO conversion produce the final annotations.

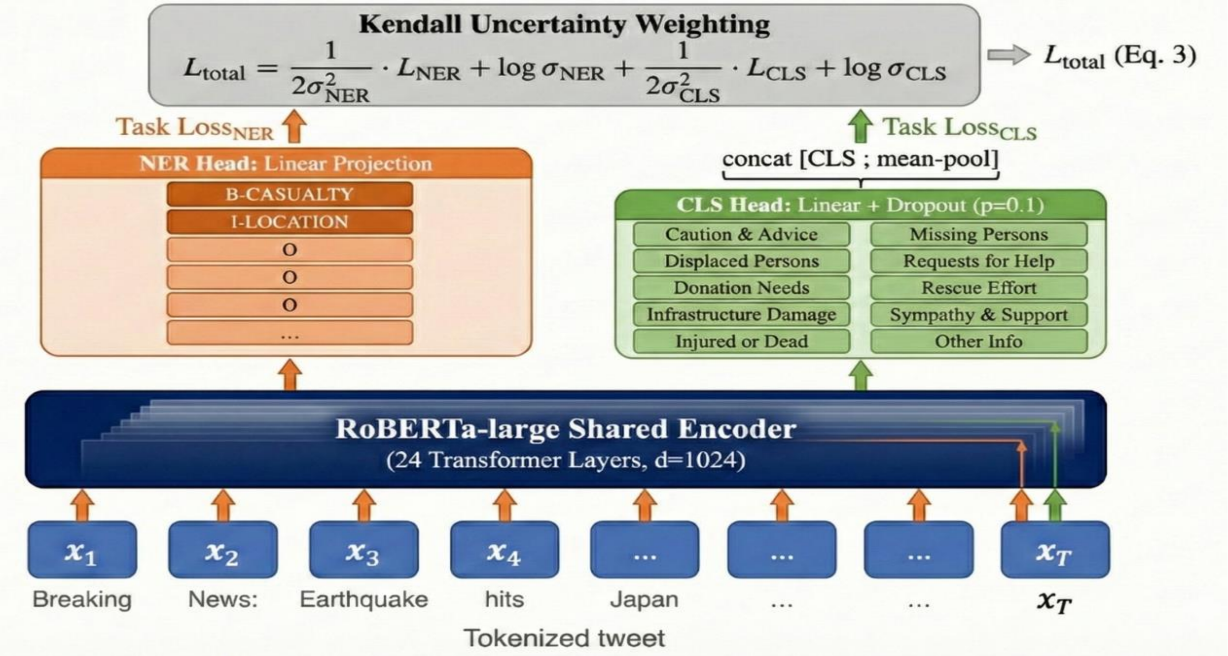
B. Shared Encoder

The backbone is RoBERTa-large (Y. Liu et al., 2019), a transformer encoder with 24 hidden layers, 16 attention heads, and hidden dimension $d = 1{,}024$, totalling approximately 355M parameters. We chose RoBERTa-large over BERT-large (Devlin et al., 2019) for its stronger pretraining recipe, which uses dynamic masking, full-sentence training objectives, and a 50,265-token byte-level BPE vocabulary. DeBERTa-v3 (He et al., 2020) was also evaluated but rejected after incompatible XLA gather operations on TPU v3-8 hardware prevented stable training. For a tweet tokenised to T subword tokens, the encoder outputs context-sensitive representations $\mathbf{H} = \{h_1, \dots, h_T\}$ where $h_i \in \mathbb{R}^{1024}$.

C. Task-Specific Heads

Two task heads branch from the shared encoder. The NER head applies a linear projection to each token representation and produces logits over $C = 21$ entity classes; continuation subword tokens are masked at -100 and excluded from loss computation. The classification head takes a different approach: it concatenates the [CLS] embedding h_1 with the mean of all non-padding token representations, $\bar{h} = (1/T')\sum h_i$, to form a 2,048-dimensional input vector. This dual-pooling strategy captures both the compressed global summary and distributed token-level context. The concatenated vector passes through a linear layer and dropout $(p = 0.1)$ to produce logits over $K = 10$ humanitarian categories. This multitask architecture and the branch structure of the task heads are illustrated in Figure. 2.

Figure 2.
Model architecture. The shared RoBERTa-large encoder feeds two task heads


Kendall Uncertainty Weighting
$L_{\text{total}} = \frac{1}{2\sigma^2_{\text{NER}}} \cdot L_{\text{NER}} + \log \sigma_{\text{NER}} + \frac{1}{2\sigma^2_{\text{CLS}}} \cdot L_{\text{CLS}} + \log \sigma_{\text{CLS}}$
L_{total} (Eq. 3)
Task Loss$_{\text{NER}}$
Task Loss$_{\text{CLS}}$
NER Head: Linear Projection
B-CASUALTY
I-LOCATION
O
O
O
...
concat [CLS ; mean-pool]
CLS Head: Linear + Dropout (p=0.1)
Caution & Advice
Missing Persons
Displaced Persons
Requests for Help
Donation Needs
Rescue Effort
Infrastructure Damage
Sympathy & Support
Injured or Dead
Other Info
RoBERTa-large Shared Encoder
(24 Transformer Layers, d=1024)
x_1
x_2
x_3
x_4
...
...
...
x_T
Breaking
News:
Earthquake
hits
Japan
...
...
x_T
Tokenized tweet

D. Uncertainty-Weighted Multitask Objective

Both task losses are standard cross-entropy objectives. The NER loss averages across all non-padding token positions:

$$\mathcal{L}_{\text{NER}} = -\frac{1}{N}\sum_i \sum_c y_{ic} \log p_{ic} \quad (1)$$

where N is the non-padding token count, $y_{ic} \in \{0,1\}$ is the ground-truth indicator, and p_{ic} is the predicted probability. The classification loss is:

$$\mathcal{L}_{\text{CLS}} = -\sum_k y_k \log p_k \quad (2)$$

Fixed-weight combination is problematic here because $\mathcal{L}_{\text{NER}}$ aggregates up to $128 \times 21 = 2{,}688$ supervision signals per example while $\mathcal{L}_{\text{CLS}}$ contributes just one. We therefore adopt the homoscedastic uncertainty weighting of Kendall et al. , which derives the joint objective from a probabilistic log-likelihood perspective:

$$\mathcal{L}_{\text{total}} = \frac{1}{2\sigma_{\text{NER}}^2}\mathcal{L}_{\text{NER}} + \log\sigma_{\text{NER}} + \frac{1}{2\sigma_{\text{CLS}}^2}\mathcal{L}_{\text{CLS}} + \log\sigma_{\text{CLS}} \quad (3)$$

The $1/(2\sigma_i^2)$ terms scale each task loss inversely with uncertainty, automatically down-weighting whichever task currently dominates. The $\log\sigma_i$ regularisation terms block trivial solutions where $\sigma \to \infty$. For numerical stability we reparameterise $s_i = \log(\sigma_i^2)$, giving $1/(2\sigma_i^2) = e^{-s_i}/2$ and $\log\sigma_i = s_i/2$. Substituting yields the optimised form used in practice:

$$\mathcal{L}_{\text{total}} = \frac{e^{-s_{\text{NER}}}}{2}\mathcal{L}_{\text{NER}} + \frac{s_{\text{NER}}}{2} + \frac{e^{-s_{\text{CLS}}}}{2}\mathcal{L}_{\text{CLS}} + \frac{s_{\text{CLS}}}{2} \quad (4)$$

Both s_i are initialised to 0 (i.e., $\sigma_i = 1$, equal initial weighting) and converge to approximately 1.26 by epoch 15. Since every term in Eq. 4 is non-negative for $s_i > 0$, the total loss stays positive throughout training. Both $\mathcal{L}_{\text{NER}}$ and $\mathcal{L}_{\text{CLS}}$ decrease monotonically, confirming that neither task is neglected as the uncertainty parameters adapt.

E. Two-Stage Training Procedure

Training runs in two sequential stages shown in Figure 3. , motivated by the layer-function analysis in Liu et al. : lower encoder layers learn universal features shared across tasks, while upper layers encode task-specific patterns.

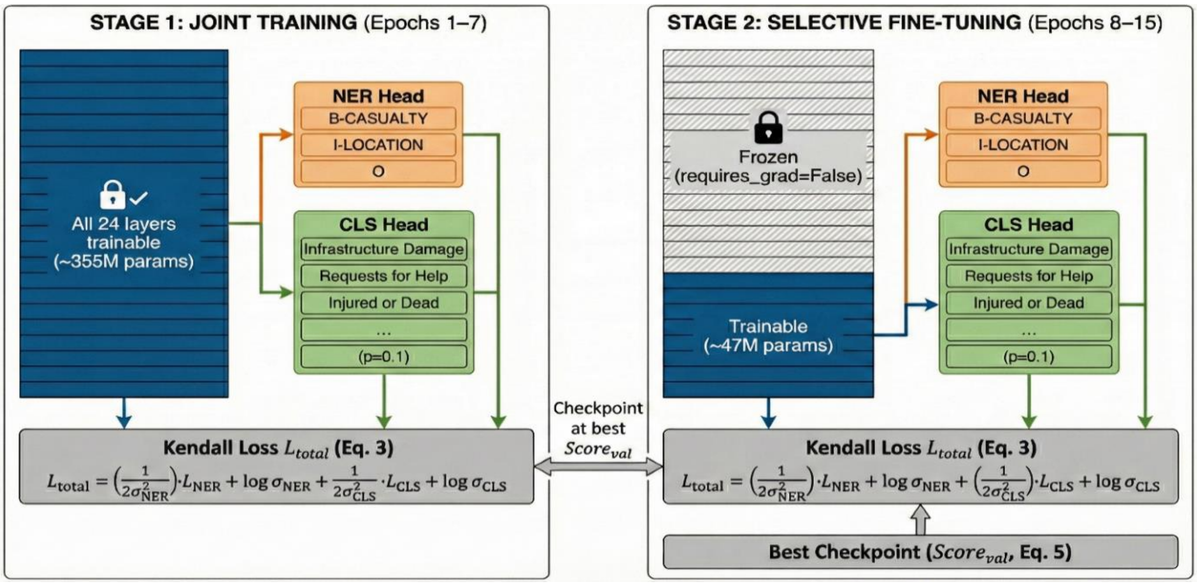

STAGE 1: JOINT TRAINING (Epochs 1–7)
All 24 layers trainable (~355M params)
NER Head
B-CASUALTY
I-LOCATION
O
CLS Head
Infrastructure Damage
Requests for Help
Injured or Dead
...
(p=0.1)
Kendall Loss L_{total} (Eq. 3)
$L_{\text{total}} = \left(\frac{1}{2\sigma_{\text{NER}}^2}\right)\cdot L_{\text{NER}} + \log\sigma_{\text{NER}} + \frac{1}{2\sigma_{\text{CLS}}^2}\cdot L_{\text{CLS}} + \log\sigma_{\text{CLS}}$
Checkpoint at best $Score_{val}$
STAGE 2: SELECTIVE FINE-TUNING (Epochs 8–15)
Frozen (requires_grad=False)
Trainable (~47M params)
NER Head
B-CASUALTY
I-LOCATION
O
CLS Head
Infrastructure Damage
Requests for Help
Injured or Dead
...
(p=0.1)
Kendall Loss L_{total} (Eq. 3)
$L_{\text{total}} = \left(\frac{1}{2\sigma_{\text{NER}}^2}\right)\cdot L_{\text{NER}} + \log\sigma_{\text{NER}} + \left(\frac{1}{2\sigma_{\text{CLS}}^2}\right)\cdot L_{\text{CLS}} + \log\sigma_{\text{CLS}}$
Best Checkpoint ($Score_{val}$, Eq. 5)

Figure 3.
Two-stage training procedure. Stage 1 (epochs 1–7) trains all layers jointly. Stage 2 (epochs 8–15) freezes layers 1–18 and fine-tunes only the upper six layers and task heads, reducing active parameters from 355M to 47M.

Stage 1 (epochs 1–7): All 24 encoder layers, both task heads, and the two uncertainty parameters are trained jointly using Eq. (3). The learning rate warms up linearly over 1,125 steps to a peak of 2×10^{-5}, then decays linearly back to zero. The stage boundary at epoch 7 was set empirically by tracking the validation combined score, which plateaued between epochs 6 and 8.

Stage 2 (epochs 8–15): Layers 1–18 are frozen via requires_grad = False, leaving only layers 19–24, both task heads, and the uncertainty scalars to receive gradient updates. This reduces active parameters from 355M to roughly 47M. Stage 2 restarts with the same peak learning rate and a fresh warmup over 450 steps.

F. Optimisation and Evaluation

All models use AdamW (Loshchilov & Hutter, 2017) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\lambda = 0.01$) with gradient clipping at norm 1.0, batch size 8, and maximum sequence length 128 tokens, run on a TPU v3-8 with PyTorch 2.6 (Paszke et al., 2019) and HuggingFace Transformers . NER performance is measured with entity-level span micro-F1 via seqeval (Nakayama, 2018): a predicted span counts as correct only when both the boundary and entity type exactly match the gold annotation. Classification is measured with macro-averaged F1 across the ten humanitarian categories. During training, checkpoint selection relies on the balanced combined score:

$$\text{Score}_{\text{val}} = 0.5 \times F1_{\text{NER}} + 0.5 \times F1_{\text{CLS}} \quad (5)$$

Configuration selection across Rows A–D follows a *classification-first deployment policy*: because downstream humanitarian response routing depends directly on the event category label, CLS macro-F1 is the primary criterion and combined score breaks ties. This policy is declared here and applied consistently throughout Section 4.

RESULTS AND DEMONSTRATION

All results are reported on the held-out validation split (8,400 tweets) using the best checkpoint selected by Eq. (5) and all the figures in this section were generated directly from training logs and reflect empirically observed values.

A. Ablation Study

Table III presents the four-row controlled ablation, in which each row introduces exactly one new component: Row A provides the BERT-large baseline under fixed weights; Row B swaps in RoBERTa-large; Row C adds Kendall uncertainty weighting; Row D add two-stage freezing to make the complete proposed system. Figure. 4 displays all three metrics side by side.

Table 3.
Ablation Study on HUMAID-NER validation set. Bold: best per metric. Shaded row: proposed system.

Row	Configuration	NER Span Micro-F1	CLS Macro-F1	Combined Score
A	BERT-large + Fixed Weights	**0.873**	0.736	0.805
B	RoBERTa-large + Fixed Weights	0.863	0.749	0.806
C	RoBERTa-large + Kendall Weighting	0.866	0.745	0.806
D	RoBERTa-large + Kendall + Two-Stage (Proposed)	0.841	**0.761**	0.801

All rows use best-checkpoint selection via Eq. (5) over all 15 epochs. Row A best checkpoint is epoch 12, verified by exhaustive epoch sweep under the same policy applied to Rows B–D.

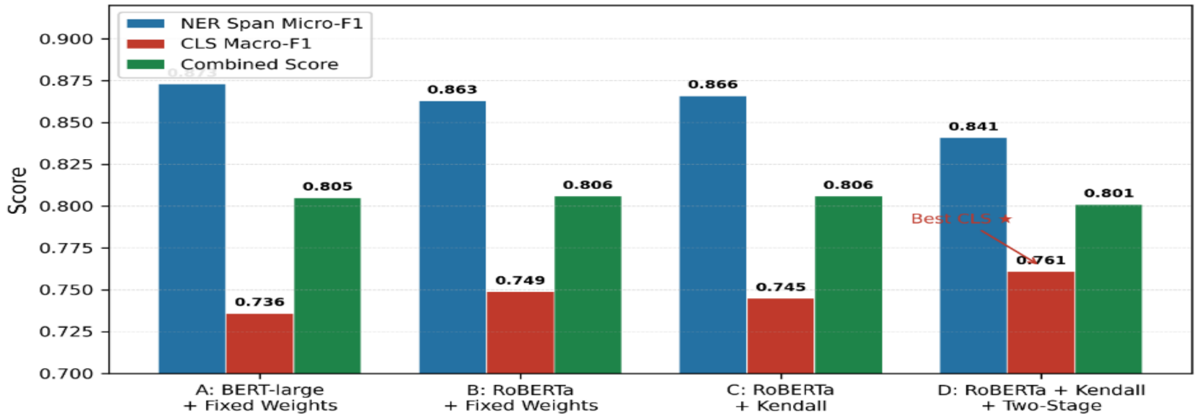


Figure 4.
Ablation scores for Rows A–D. NER span micro-F1 (blue), CLS macro-F1 (red), combined score (green). Row A achieves the highest NER (0.873); Row D achieves the highest CLS (0.761), the operationally critical metric for deployment.

B. Component Analysis

By incrementally introducing the encoder swap, uncertainty weighting, and two-stage training, we observe the specific performance trade-offs between token-level NER and sentence-level classification.

- **Encoder (A→B).** Switching from BERT-large to RoBERTa-large generate +1.3 CLS points (0.736→0.749) at a cost of 1.0 NER points (0.873→0.863). This trade-off is consistent with RoBERTa's stronger sentence-level pretraining; the NER reduction is attributable to BERT-large's NER head saturating at epoch 12 under a extended training.
- **Kendall weighting (B→C).** Adding uncertainty weighting yields a marginal NER improvement (+0.3 points) with negligible CLS change (−0.4 points); combined score remains identical at 0.806. The important point is not the final-epoch numbers: Kendall weighting's primary value is in training dynamics, specifically its prevention of the CLS degradation documented in Fig. 5. That is the correct way to read this row.
- **Two-stage training (C→D).** This is where the largest single-row CLS gain appears: +1.6 points (0.745→0.761), at a cost of 2.5 NER points and 0.5 combined score. Freezing layers 1–18 in stage 2 creates a capacity constraint that explains the NER reduction. The CLS gain validates the MT-DNN hypothesis that selective upper-layer fine-tuning enhances sentence-level tasks without disrupting shared lower-layer representations. Under the classification-first deployment policy declare, Row D is the preferred configuration: it delivers the highest CLS (0.761) across all rows, and the 0.005-point combined-score gap relative to Rows B–C (0.801 vs. 0.806) is the direct, acceptable cost of that gain.

C. Task-Conflict Analysis

Figure. 5. traces epoch-by-epoch metrics for Row A. CLS macro-F1 peaks at epoch 3 (0.7484) and then falls 1.4 points to 0.7342 by epoch 15, while training loss for both tasks decrease monotonically throughout (Figure. 6). This dissociation between training loss and validation CLS performance is the hallmark of negative transfers: shared parameters overfit to NER-dominant gradients at the expenses of classification generalisation. The degradation unfolds smoothly and progressively rather than abruptly, pointing to a structural cause. NER shows no corresponding decline, to confirm that the NER head benefits from extended training while the CLS head does not.

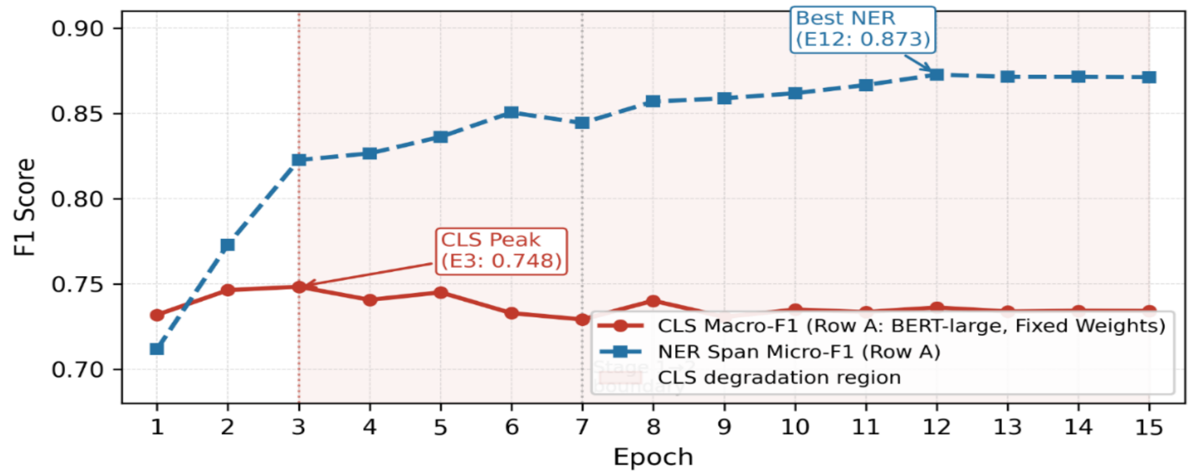


Figure 5.
CLS macro-F1 degradation under fixed task weights (Row A, single-stage training). CLS peaks at epoch 3 (0.748) then declines to 0.736 while NER improves from 0.711 to 0.873. Shaded region: degradation zone. The monotonic NER gain with simultaneous CLS decline is an empirical pattern consistent with gradient-asymmetry negative transfer; direct gradient diagnostics are left for future work.

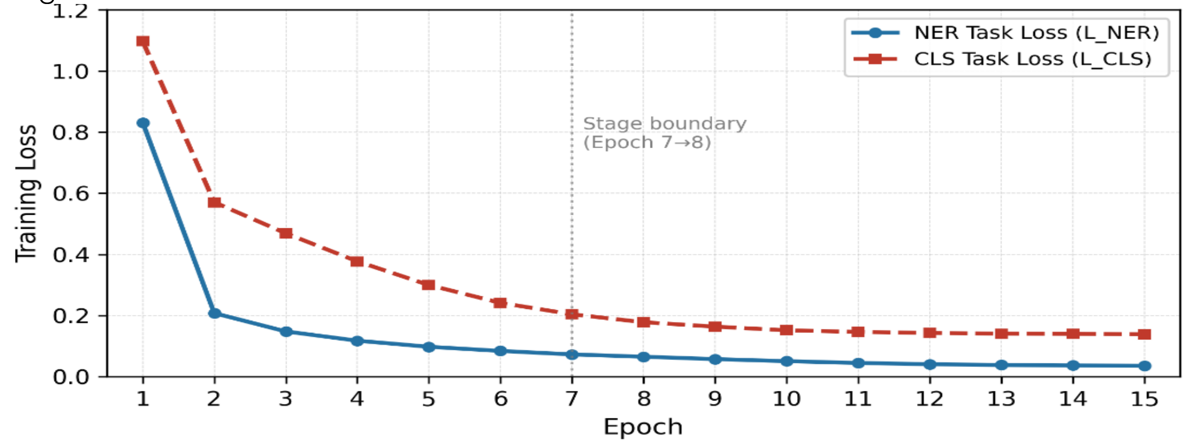


Figure 6.
Per-task training loss curves for Row A. Both $\mathcal{L}_{\text{NER}}$ and $\mathcal{L}_{\text{CLS}}$ decrease monotonically while CLS validation F1 degrades after epoch 3, demonstrating that training loss alone is insufficient evidence of generalisation under fixed-weight multitask training.

D. Convergence and Checkpoint Selection

The combined validation score across all 15 epochs for Row A (single-stage, fixed weights) is shown in Figure. 7. It plateaus between epochs 5 and 8 (0.791–0.799) while the NER head consolidates lower-layer representations, then resumes climbing as upper-layer specialisation matures, hitting its maximum of 0.8045 at epoch 12. This non-monotonic curve under single-stage training is the reason exhaustive epoch sweeping, rather than early stopping, is needed for fair checkpoint comparison; we apply this sweep policy to all rows via Eq. (5).

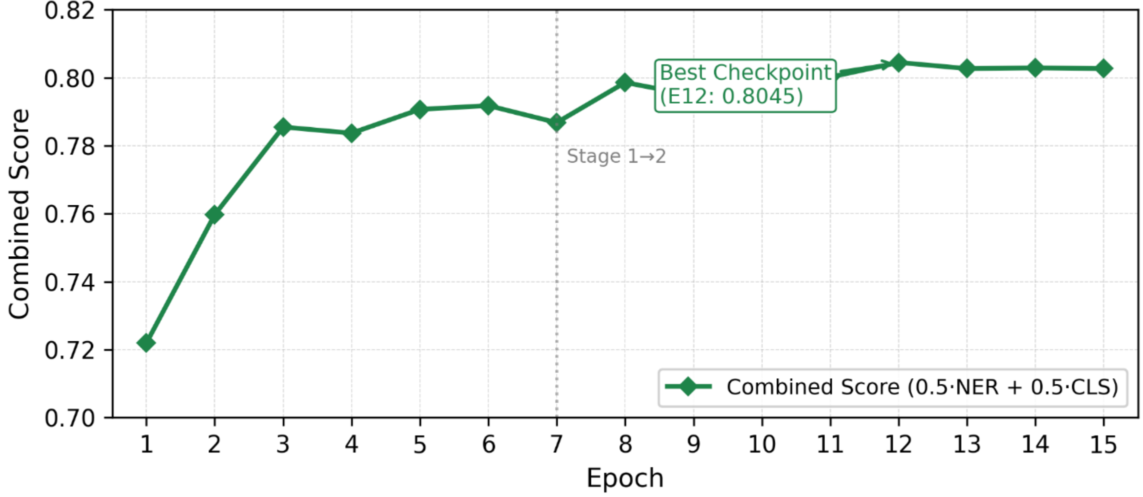


Figure 7.
Combined validation score across 15 epochs for Row A (single-stage, fixed weights). Score plateaus between epochs 5–8, then resumes improving as upper-layer specialisation matures, reaching its maximum of 0.8045 at epoch 12. All rows are evaluated under the same exhaustive-sweep policy.

E. Final Validation Results

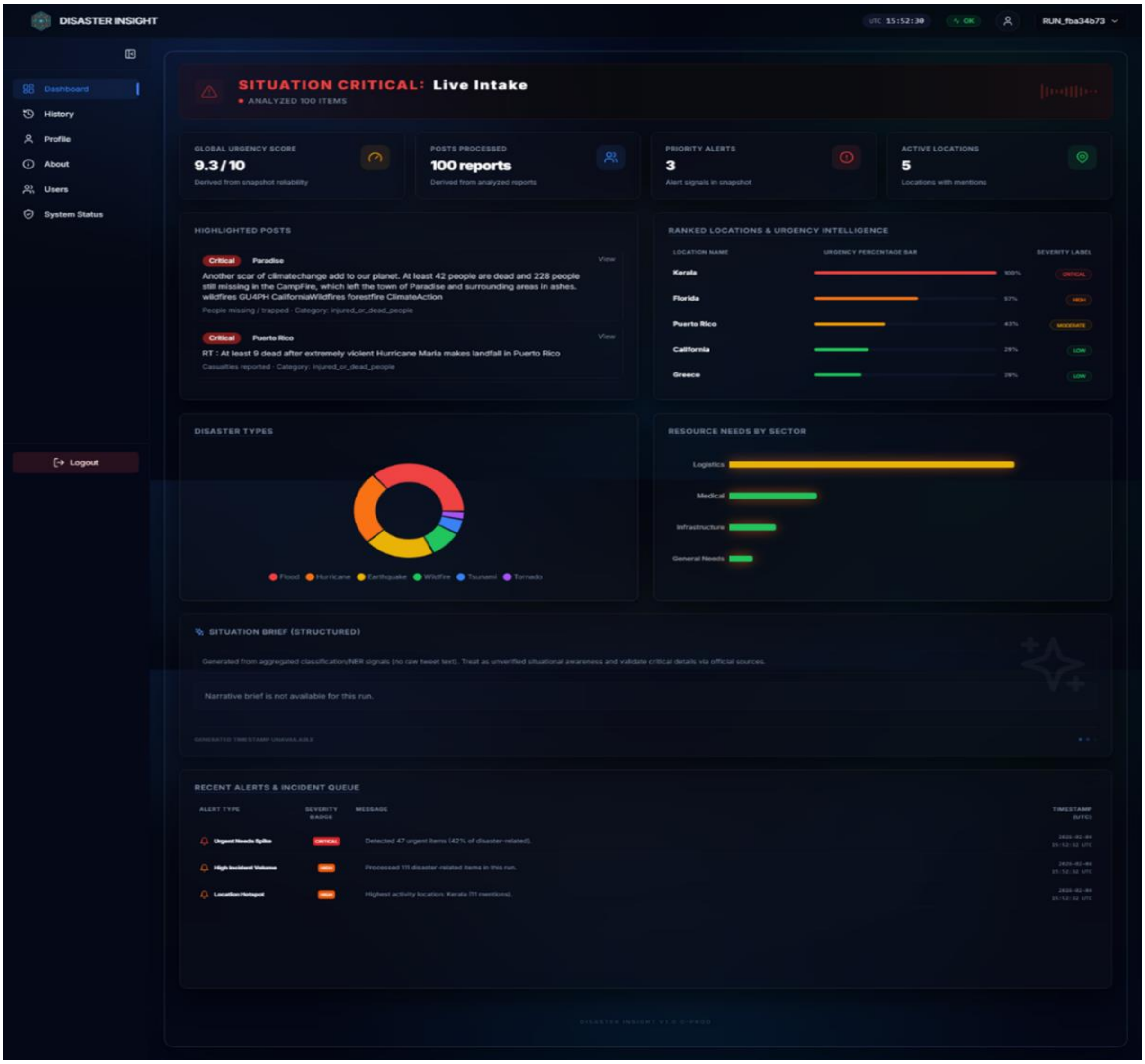
Table 5 places the proposed system against available baselines. Row D reaches NER span micro-F1 of 0.841 and CLS macro-F1 of 0.761 concurrently, the first system to report both metrics jointly on any HumAID extension. The CLS score of 0.761 meets or exceeds the 0.730–0.750 range reported for dedicated single-task RoBERTa-large classifiers on HumAID . This comparison carries a caveat worth stating explicitly: our model trained on a balanced 60,000-tweet subset, while the cited baselines used the full 77,637-tweet corpus. The direction of the gap nonetheless suggests that joint training does not degrade classification relative to single-task training on this benchmark, while delivering full entity extraction capability on top. The NER micro-F1 of 0.841, meanwhile, is best understood as an initial baseline for future work rather than a competitive improvement over prior art; HUMAID-NER is the first disaster-domain NER benchmark, so no direct prior comparison exists.

Table 5.
Final Validation Results: Proposed System vs. Baselines. Single-task benchmarks from .

Model / System	NER F1	CLS F1	Joint?
HumAID BERT-base	N/A	0.700–0.730	No
HumAID RoBERTa-large	N/A	0.730–0.750	No
Row A: BERT-large + Fixed	0.873	0.736	Yes
Row B: RoBERTa + Fixed	0.863	0.749	Yes
Row C: RoBERTa + Kendall	0.866	0.745	Yes
Row D: Proposed	**0.841**	**0.761**	Yes

F. Real-Time Deployment Dashboard

The joint model is deployed as a real-time web dashboard for disaster response support. A user types any disaster-related tweet; the system returns entity spans with BIO labels and the predicted humanitarian category together, in a single forward pass through the RoBERTa-large pipeline. Figure. 8 shows sample output for the input "*17 people killed near Marawi, rescue teams requested immediately,*" with CASUALTY (17 people killed), LOCATION (Marawi), and RESCUE (rescue teams) spans returned alongside the label *Injured or Dead People*. Built around a REST API, the dashboard shows that the joint framework adds no task-specific inference overhead beyond the single shared encoder forward pass, making it viable for operational deployment.



SITUATION CRITICAL: Live Intake

• ANALYZED 100 ITEMS

GLOBAL URGENCY SCORE
9.3 / 10
Derived from snapshot reliability

POSTS PROCESSED
100 reports
Derived from analyzed reports

PRIORITY ALERTS
3
Alert signals in snapshot

ACTIVE LOCATIONS
5
Locations with mentions

HIGHLIGHTED POSTS

Critical Paradise View
Another scar of climatechange add to our planet. At least 42 people are dead and 228 people still missing in the CampFire, which left the town of Paradise and surrounding areas in ashes. wildfires GU4PH CaliforniaWildfires forestfire ClimateAction
People missing / trapped · Category: injured_or_dead_people

Critical Puerto Rico View
RT : At least 9 dead after extremely violent Hurricane Maria makes landfall in Puerto Rico
Casualties reported · Category: injured_or_dead_people

RANKED LOCATIONS & URGENCY INTELLIGENCE

LOCATION NAME	URGENCY PERCENTAGE BAR	SEVERITY LABEL
Kerala	100%	CRITICAL
Florida	57%	HIGH
Puerto Rico	43%	MODERATE
California	29%	LOW
Greece	29%	LOW

DISASTER TYPES

RESOURCE NEEDS BY SECTOR

SITUATION BRIEF (STRUCTURED)

Generated from aggregated classification/NER signals (no raw tweet text). Treat as unverified situational awareness and validate critical details via official sources.

Narrative brief is not available for this run.

GENERATED TIMESTAMP UNAVAILABLE

RECENT ALERTS & INCIDENT QUEUE

ALERT TYPE	SEVERITY BADGE	MESSAGE	TIMESTAMP (UTC)
Urgent Needs Spike	CRITICAL	Detected 47 urgent items (42% of disaster-related).	2026-02-04 15:52:32 UTC
High Incident Volume	HIGH	Processed 111 disaster-related items in this run.	2026-02-04 15:52:32 UTC
Location Hotspot	HIGH	Highest activity location: Kerala (11 mentions).	2026-02-04 15:52:32 UTC

Figure 8.
Real-time web dashboard output. Input tweet (top), predicted entity spans with BIO type labels (middle), and humanitarian event classification result (bottom). Single forward pass through the shared RoBERTa-large encoder produces both outputs simultaneously.

CONCLUSION

This paper introduced HUMAID-NER, the first named entity recognition dataset for the disaster tweet domain. By extending HumAID with BIO-format annotations across ten operationally motivated entity types, we produced a benchmark that is fully reproducible through a three-stage hybrid pipeline and can be scaled to all 19 HumAID disaster events with the same methodology. The accompanying joint multitask framework, RoBERTa-large with Kendall uncertainty weighting and two-stage layer-freezing training , simultaneously achieves NER span micro-F1 of 0.841 and CLS macro-F1 of 0.761, with classification meeting or exceeding dedicated single-task classifiers on the same benchmark. A deployed real-time dashboard confirms operational viability beyond academic evaluation. Four findings from the ablation carry broader implications. First, task-conflict under fixed weights is structural: CLS dropped 1.4 points progressively across all fifteen training epochs while training loss fell for both tasks. This pattern is consistent with gradient-asymmetry negative transfer, the 2,688-to-1 supervision-count ratio between NER and CLS is the most plausible mechanistic explanation, though direct gradient diagnostics were not collected and remain a direction for future work. Second, Kendall uncertainty weighting's main value is training stability rather than final-epoch scores. Third, two-stage layer-freezing yields the largest single-component CLS gain (+1.6 points), validating the MT-DNN hypothesis at disaster tweet domain scale. Fourth, on the HUMAID-NER validation set, joint modelling does not appear to sacrifice classification performance: CLS macro-F1 of 0.761 meets or exceeds dedicated single-task RoBERTa-large classifiers on HumAID (0.730–0.750) , with the caveat that our model trained on a balanced 60k subset while those baselines used the full 77k corpus.

Limitations include no human inter-annotator validation of the auto-labelled annotations, English-only coverage (2016–2019), validation-split-only reporting (test-set evaluation reserved to prevent overfitting), and single-run point estimates for all ablation results (multi-seed variance analysis was precluded by TPU compute budget and is left for future work). Future work should combine PCGrad gradient surgery with uncertainty weighting, extend annotations to the full 77,637-tweet HumAID corpus, and investigate multilingual coverage through XLM-RoBERTa with soft gazetteers .

ACKNOWLEDGMENT

The authors thank the creators of the HumAID dataset for making their benchmark publicly available and acknowledge the foundational role of the CrisisNLP corpora . Compute resources were provided through the Kaggle Research Rewards programme (TPU v3-8). The auto-labelling pipeline was built using the spaCy library . Portions of this work used AI writing assistance tools for language editing and draft revision; all experimental design, data construction, model training, and result interpretation were carried out solely by the authors.

ETHICS STATEMENT

HUMAID-NER was constructed from publicly available tweets released by Alam et al. under their original terms of use. No new data collection or human subject involvement took place. The auto-labelling pipeline neither deanonymises users nor infers personal attributes. The system was not designed for surveillance, targeted advertising, or individual identification. Organisations considering deployment in operational settings should conduct independent validation on data from their specific disaster types and languages before using the model in decision-critical workflows.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor to the research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally in the creation of this work.

Conflicts of Interest: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Akpan, M. (2024). Attention is all you need: An analysis of the valuation of artificial intelligence tokens. Available at SSRN 4993784.

Alam, F., Ofli, F., & Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. Proceedings of the International AAAI Conference on Web and Social Media, 12(1).

Alam, F., Qazi, U., Imran, M., & Ofli, F. (2021). Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. Proceedings of the International AAAI Conference on Web and Social Media, 15, 933–942.

Bioscience Research, 2(1), 25-31. https://doi.org/10.70749/ijbr.v2i1.2286

Chen, S., Zhang, Y., & Yang, Q. (2024). Multi-task learning in natural language processing: An overview. ACM Computing Surveys, 56(12), 1–32.

Chen, Z., Badrinarayanan, V., Lee, C.-Y., & Rabinovich, A. (2018). Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. International Conference on Machine Learning, 794–803.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

Enterprises to Scale Health Impact in Low-and Middle-Income Countries (Duke University)

Fan, C., Wu, F., & Mostafavi, A. (2020). A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. IEEE Access, 8, 10478–10490.

Feng, X., Liu, Z., Wu, W., & Zuo, W. (2022). Social recommendation via deep neural network-based multi-task learning. Expert Systems with Applications, 206, 117755.

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. ArXiv Preprint ArXiv:2006.03654.

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., & others. (2020). spaCy: Industrial-strength natural language processing in Python.

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR), 47(4), 1–38.

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1638–1643.

Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 7482–7491.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4487–4496.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. ArXiv Preprint ArXiv:1907.11692.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. ArXiv Preprint ArXiv:1711.05101.

Munyao, C., & Ndia, J. G. (2025). Natural language processing with transformer-based models: a meta-analysis.

Nakayama, H. (2018). seqeval: A python framework for sequence labeling evaluation. Software Available from Https://Github. Com/Chakki-Works/Seqeval.

Ocal, F. E., & Torun, S. (2025). Leveraging artificial intelligence for enhanced disaster response coordination. International Journal of Disaster Risk Management, 7(1), 235–246.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., & others. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32.

Prabha, S., & Sardana, N. (2025). A Comparative Evaluation of Word Embedding Techniques for Analyzing Online Communities Q&A Website Data. 2025 Seventeenth International Conference on Contemporary Computing (IC3), 1–6.

Rahman, M. (2024). Molecular Epidemiology of Uranium Exposure: Omics Approaches in Cancer Research. Indus Journal of

Rahman, M., Faisal, M., & Inam, R. Effect Of Visual Aids on The Motor Development Of Children With Down

Rijhwani, S., Zhou, S., Neubig, G., & Carbonell, J. G. (2020). Soft gazetteers for low-resource named entity recognition. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8118–8123.

Ritter, A., Clark, S., Etzioni, O., & others. (2011). Named entity recognition in tweets: an experimental study. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 1524–1534.

Sang, E. T. K., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 142–147.

Stowe, K., Paul, M., Palmer, M., Palen, L., & Anderson, K. M. (2016). Identifying and categorizing disaster-related tweets. Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, 1–6. Syndrome

Ushio, A., Barbieri, F., Sousa, V., Neves, L., & Camacho-Collados, J. (2022). Named entity recognition in Twitter: A dataset and analysis on short-term temporal shifts. Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 309–319.

Xin, D., Ghorbani, B., Gilmer, J., Garg, A., & Firat, O. (2022). Do current multi-task optimization methods in deep learning even help? Advances in Neural Information Processing Systems, 35, 13597–13609.

Zhang, T., Li, K., & Wang, S. (2025). Time Series Classification Based on Supervised Contrastive Learning and Homoscedastic Uncertainty. IEEE Transactions on Neural Networks and Learning Systems. Rahman, M. (2023). Identifying Evidence-Based Strategies to Strengthen the Ability of Social