applied sciences

MDPI

Article

Fine-Grained Semantic Classification of Disaster-Related Social Media Text for Emergency Management

Xiaodong Wang [1] and Tengfei Yang [2,*]

[1] School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang 471023, China; 9906208@haust.edu.cn
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: yangtf@radi.ac.cn

Abstract

Disaster-related social media posts often report casualties, rescue needs, infrastructure damage, shelter demand, and local situation changes earlier than formal channels, yet their brevity and noise make operational classification difficult. This study examines whether a practical and reproducible classification pipeline can support fine-grained, emergency-oriented semantic recognition under a deliberately conservative evaluation setting. We convert 14,392 English tweets from CrisisSense-LLM into six actionable semantic categories, partition the data by proxy event groups, and compare a TF-IDF logistic-regression baseline, supervised BERT-base fine-tuning, and zero-shot natural language inference. The evaluation is further extended to mapped HumAID data, a manually reviewed 177-post Chinese boundary-test set, a Chinese-to-English translation bridge, and a fixed-budget selective adjudication simulation. BERT-base achieves the best held-out main-test performance (Macro-F1 = 0.8824), outperforming TF-IDF (0.6133) and zero-shot inference (0.3581), and reaches 0.8132 Macro-F1 on HumAID without retraining. Direct English-to-Chinese transfer is ineffective, whereas multilingual BERT and translation bridging improve Chinese Macro-F1 to 0.2684 and 0.3603, respectively. With 600 reviewed posts, selective adjudication reaches 0.7792 Macro-F1 on the main test set and 0.7153 on HumAID. These findings indicate that the central contribution is not a new model architecture, but an empirically validated workflow that combines supervised fine-tuning, leakage-aware evaluation, external validation, cross-lingual stress testing, and information-driven human review. The novelty therefore lies in the reproducible integration of data mapping, group-aware evaluation, external and cross-lingual stress testing, and selective human review into a single emergency-oriented assessment workflow.

Keywords: disaster social media; fine-grained semantic classification; emergency management; cross-event generalization; cross-lingual transfer; human-in-the-loop learning

Academic Editor: Douglas O'Shaughnessy

Received: 11 May 2026
Revised: 24 May 2026
Accepted: 3 June 2026
Published: 4 June 2026

1. Introduction

Social media platforms have become important sources of rapidly emerging disaster information. During emergencies, short posts often report casualties, requests for help, rescue activities, shelter needs, infrastructure damage, and local situation updates earlier than many formal channels. At the same time, this information stream is noisy, incomplete, and highly uneven across events. In the converted CrisisSense-LLM corpus used here, 4026 of 14,392 posts (28.0%) lack a usable location field, the largest proxy event group contains 1936 posts, and the smallest target class, request for help, accounts for only

824 posts (5.7%). A single disaster may generate many near-duplicate posts containing the same place names and recurring expressions, whereas other events may be represented by only a small number of messages. For emergency management, the core challenge is therefore not merely to detect whether a post is disaster-related, but to identify which actionable semantic category it conveys and whether the classifier remains reliable across events, datasets, and languages [1–3].

Recent studies have demonstrated the value of disaster-related social media classification, but three limitations remain prominent. First, many studies still emphasize relevance detection or coarse humanitarian labeling, whereas emergency decision making often requires finer distinctions among casualties, help requests, rescue activity, shelter or supply demand, infrastructure damage, and general situational updates [2–4]. Second, text-level random splitting can yield optimistic estimates because posts from the same event frequently share location names, recurring entities, and highly similar expressions [5–7]. Third, low-resource settings such as Chinese disaster social media remain underexplored, and cross-lingual transfer is often reported without a clearly bounded evaluation design [8–12].

These limitations motivate a pipeline-oriented research question: can a relatively standard supervised classifier become useful for emergency text triage if the task definition, evaluation split, external validation, cross-lingual testing, and human review strategy are designed carefully? This framing is important because operational risk does not only come from a weak classifier. It can also come from an overly broad label space, leakage between training and test events, failure on another dataset, failure under language shift, or inefficient use of scarce human review time.

Accordingly, this study builds a complete empirical pipeline rather than proposing another complex architecture. The main task is framed as a six-way fine-grained semantic discrimination of English disaster tweets. The evaluation protocol is tightened through a proxy event-group split. The resulting model is then tested on an external English benchmark and on a small manually reviewed Chinese social media set. Finally, the study examines whether a selective adjudication strategy can use limited human review effort more effectively than random sample addition.

The empirical analysis is organized around four questions. First, how much performance is gained by supervised contextual modeling compared with sparse lexical features and zero-shot inference? Second, which semantic categories remain difficult after fine-tuning? Third, does the trained model retain useful performance on an external English disaster dataset and under Chinese boundary testing? Fourth, when only a limited number of posts can be reviewed, does selecting high-information cases improve the model more efficiently than random review?

For clarity, the workflow does not claim architectural novelty in TF-IDF, BERT, BART-MNLI, or mBERT. Its contribution is instead a task-to-evaluation design that makes these standard components operationally comparable under one fine-grained emergency label space, one leakage-aware split, one external English validation, one Chinese boundary test, and one fixed-budget human review simulation.

The main contributions of this study are as follows:

1. It establishes a unified research framework that combines fine-grained disaster semantic classification, cross-dataset validation, cross-lingual boundary testing, and budget-constrained human-in-the-loop updating.
2. It replaces random text-level splitting with a proxy event-group split, thereby providing a more conservative evaluation setting for cross-event generalization.
3. It validates the main model on the HumAID benchmark and on a manually reviewed Chinese social media set, allowing external and cross-lingual generalization to be discussed separately.

4. It demonstrates that selective adjudication, which combines model uncertainty and model disagreement, improves label efficiency over random adjudication under a fixed human review budget [13–15].

2. Related Work

2.1. Disaster Social Media Classification

Disaster social media analysis has been a central topic in crisis informatics for more than a decade [1,5,16]. A broad survey by Imran et al. showed that filtering, classification, extraction, and summarization are recurring computational tasks in emergency-related social media processing [1]. Within this area, HumAID has become a widely used benchmark for humanitarian tweet classification and provides large-scale, human-annotated disaster data across multiple events [2]. More recently, CrisisSense-LLM extended this line of work toward finer-grained disaster semantics by organizing posts into a more detailed labeling setting suitable for actionable information identification [3]. Xie et al. further showed that richer contextual modeling improves disaster text classification beyond shallow lexical baselines [4]. Together, these studies indicate a shift from coarse crisis relevance detection toward finer semantic understanding [5,16].

2.2. Zero-Shot and Large-Model Approaches

The growing use of large language models has encouraged researchers to examine whether zero-shot or instruction-based approaches can reduce annotation costs in crisis informatics. CrisisSense-LLM itself illustrates the promise of task-oriented instruction tuning for disaster-related social media analysis [3]. However, current evidence does not support the assumption that general zero-shot inference is sufficient for reliable operational classification. McDaniel et al. reported substantial variability in zero-shot crisis tweet classification, with performance remaining sensitive to prompt design and dataset conditions [8]. For this reason, zero-shot inference is better treated as a meaningful comparison point than as a replacement for supervised modeling in short, noisy disaster text [8,17,18].

2.3. Human-in-the-Loop Learning and Cross-Lingual Transfer

Human review remains critical in disaster scenarios because data distributions change quickly and annotation resources are scarce [13,14]. Pohl et al. demonstrated that active learning can improve the efficiency of social media labeling for crisis management [19]. Kaufhold et al. further showed that active, incremental, and online learning can support rapid relevance classification in disaster settings [20]. Pandey et al. emphasized that human-in-the-loop systems depend not only on whether humans participate, but also on how review effort is scheduled and how annotation errors are controlled [21].

Cross-lingual transfer is particularly relevant for low-resource disaster communication settings [11,12]. Pires et al. showed that multilingual BERT exhibits non-trivial zero-shot transfer, although its robustness varies across languages and scripts [9]. Conneau et al. later demonstrated that large-scale multilingual pretraining can substantially strengthen cross-lingual representation learning [10]. These findings suggest that if an English-only model fails on Chinese disaster text while a multilingual model partially recovers performance, this pattern is theoretically expected rather than incidental [9–12].

Compared with these strands of work, the present study is positioned less as a model-invention paper and more as an evaluation-and-deployment study. Prior benchmarks, zero-shot methods, active learning, and cross-lingual representation learning each address one part of the problem. The gap is that emergency use requires these parts to be examined together: fine-grained labels must be tested under a conservative split, external validation

must be separated from cross-lingual transfer, and human review should be evaluated as a limited resource rather than as an unlimited annotation assumption.

3. Materials and Methods

3.1. Study Overview

The method is organized as a four-stage pipeline: data preparation, fine-grained semantic classification, out-of-domain validation, and selective adjudication. The first stage produces a six-category disaster text dataset. The second stage compares three representative classification strategies. The third stage evaluates generalization to both an external English benchmark and a Chinese microblog set. The fourth stage simulates a budget-constrained human-in-the-loop update process. Figure 1 summarizes this workflow.

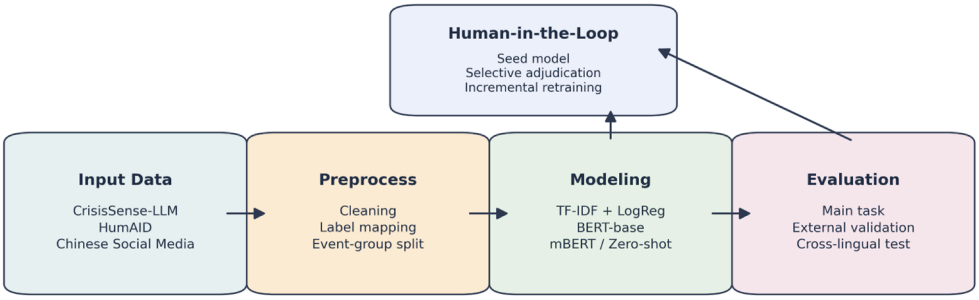

Figure 1. Study framework and experimental workflow.

Disaster-related posts are cleaned and mapped into six semantic categories, then used for main-task training, external validation, cross-lingual testing, and selective adjudication experiments.

Figure 1 presents the overall study framework and experimental workflow. The study first converts disaster-related social media posts into a six-category fine-grained semantic label space, and partitions the main dataset using a proxy event-group split to reduce potential train-test leakage. Based on this conservative evaluation protocol, three classification strategies, namely TF-IDF with logistic regression, supervised BERT-base fine-tuning, and zero-shot BART-MNLI, are compared on the held-out main test set. The supervised BERT-base classifier is then further examined in two out-of-domain settings: external validation on the mapped HumAID benchmark, and cross-lingual boundary testing on a manually reviewed Chinese social media set. Finally, a selective adjudication simulation evaluates whether limited human review can be allocated more efficiently by prioritizing samples with high uncertainty and model disagreement. In this way, the framework connects data preparation, model comparison, external generalization, cross-lingual robustness, and human-in-the-loop updating into a unified evaluation pipeline.

3.2. Data Sources and Label Space

The main dataset consists of 14,392 English disaster tweets converted from CrisisSense-LLM [3]. To maintain label stability and direct comparability across experiments, the task is defined using six semantic categories: casualty, infrastructure damage, request for help, rescue or aid, shelter or supply, and situation update. These categories were selected because each corresponds to a type of information that can plausibly support emergency triage. Casualty refers to reports of deaths, injuries, or missing persons; infrastructure damage refers to physical damage to roads, bridges, buildings, utilities, or other facilities; request for help captures explicit or implicit unmet needs; rescue or aid describes ongoing or completed assistance; shelter or supply covers evacuation, accommodation, food, water, medicine, and other relief needs; situation update captures general

descriptions of local conditions when no other specific operational category dominates. Although the source literature often frames disaster post understanding as a multi-label problem, the converted data used here behave predominantly as single-primary-label examples. Accordingly, the present study is more accurately described as fine-grained semantic discrimination than as full joint multi-label prediction.

Table 1 summarizes the target categories, operational annotation rules, source-label mappings, and mapped counts.

Table 1. Target categories, annotation rules, source-label mapping, and mapped counts.

Target Category	Operational Annotation Rule	CrisisSense-LLM Source Label(s), Count	HumAID Source Label(s), Test-Count
casualty	Deaths, injuries, missing persons, trapped persons, or victim counts	injured_or_dead_people + missing_or_found_people, 2745	injured_or_dead_people + missing_or_found_people, 1519
infrastructure_damage	Damage to buildings, roads, bridges, utilities, communications, or other facilities	infrastructure_and_ utility_damage, 2051	infrastructure_and_ utility_damage, 1617
request_for_help	Explicit or implicit requests for rescue, transfer, assistance, or urgently needed materials	requests_or_urgent_needs, 824	requests_or_urgent_needs, 521
rescue_or_aid	Rescue operations, official response, volunteering, donation, transport, or aid distribution	rescue_volunteering_or_ donation_effort, 6290	rescue_volunteering_or_ donation_effort, 4219
shelter_or_supply	Evacuation, shelter, food, water, medicine, tents, or other relief supplies	displaced_people_and_ evacuations, 975	displaced_people_and_ evacuations, 790
situation_update	General disaster-status updates, warnings, advice, official notices, or broader relevant information	caution_and_advice, 1507	caution_and_advice + other_relevant_information, 3477

For HumAID, not_humanitarian, sympathy_and_support, and dont_know_cant_judge were excluded before external validation because they do not correspond to any of the six operational categories. The converted CrisisSense-LLM main corpus contains 14,392 retained domain rows, each with one mapped primary semantic label; the mapped HumAID test split contains 12,143 rows after the same six-category filtering.

External validation is conducted on the mapped HumAID test split, which contains 12,143 English disaster tweets after label alignment and category filtering [2]. Cross-lingual testing is conducted on a manually reviewed Chinese social media set containing 177 posts. The set contains 160 informative posts with at least one semantic label and 17 non-informative or out-of-scope posts encoded with no positive semantic label. This Chinese set is intentionally treated as a boundary test rather than as a full benchmark or a training resource. Its purpose is to expose whether an English-trained system collapses when faced with Chinese disaster posts, not to support a final claim about Chinese deployment performance.

The Chinese set was constructed as a manually reviewed boundary-test set rather than a training benchmark. It combines 120 previously locked gold rows with 57 additional rows selected from a candidate sheet using a six-row event cap and priority flags for rare labels, new events, and high-ambiguity cases. The final set contains 145 Weibo posts,

23 Bilibili comments, and nine Zhihu items across 17 event groups. Its gold labels contain 105 multi-label informative rows, 55 single-label informative rows, and 17 non-informative rows. The label counts are: rescue_or_aid = 130, shelter_or_supply = 69, situation_update = 65, casualty = 29, request_for_help = 17, and infrastructure_damage = 13. Annotation followed an informative-first rule: non-informative posts receive no semantic label, while informative posts are assigned one or more of the six operational categories. Negated statements such as no casualties or no building collapse are treated as situation updates rather than casualty or infrastructure-damage labels. A formal inter-annotator agreement score was not computed; this is why the set is used only as a boundary test and not as a definitive Chinese benchmark.

3.3. Split Protocol and Evaluation Metrics

Instead of using a random text-level split, the main dataset is partitioned through a proxy event-group strategy based on disaster type and primary location. This design aims to reduce leakage caused by highly similar posts from the same event appearing in both training and test data. For example, posts associated with the same normalized disaster type and primary location are assigned to the same split so that the test set is less likely to contain near-duplicates of training examples. Under this protocol, the main corpus contains 10,074 training posts, 1439 validation posts, and 2879 test posts. The validation split is used for checkpoint selection and model selection, while the test split is held out for final reporting only. The proxy event group is constructed as normalized disaster type:normalized primary location. Disaster type is read from the source event category and normalized; for example, wildfires to wildfire and floods to flood. Primary location is the first usable entry in the source location list after removing blanks and the placeholder yyy; if no usable location remains, the location field is set to unknown. This procedure produces 1658 proxy event groups: 563 groups in training, 543 in validation, and 552 in test. Row counts are 10,074, 1439, and 2879, respectively. The split assignment sorts groups by size and assigns whole groups to train/dev/test targets of approximately 70%/10%/20%, so no proxy event group is shared across splits.

Evaluation is reported using Subset Accuracy, Macro-F1, Micro-F1, and per-class F1. In the present setting, Subset Accuracy is effectively equivalent to exact sample-level classification accuracy because the data are dominated by one primary label per post. Macro-F1 is treated as the principal indicator because it better reflects performance across infrequent and frequent classes. All metrics are computed on six-dimensional binary label vectors. Subset Accuracy is the proportion of posts for which the complete predicted label vector exactly matches the gold vector. Macro-F1 is the unweighted mean of the six per-label F1 values. Micro-F1 pools true positives, false positives, and false negatives over all six labels before computing F1. Therefore, even when the main gold labels are single-primary-label examples, Subset Accuracy and Micro-F1 need not be identical: an empty prediction or an extra positive label changes the exact-match score, and also changes the label-level false-positive or false-negative counts.

3.4. Compared Models

The model comparison is designed to separate three practical questions: whether shallow lexical features are sufficient, whether supervised contextual fine-tuning provides a stronger operational model, and whether zero-shot inference can replace supervised training. Three model families are therefore compared on the main task:

A sparse-feature baseline using term frequency-inverse document frequency (TF-IDF) and logistic regression.

A supervised Bidirectional Encoder Representations from Transformers base model (BERT-base) classifier used as the main model [22].

A zero-shot natural language inference (NLI) baseline based on a Bidirectional and Auto-Regressive Transformers (BART) model fine-tuned on Multi-Genre Natural Language Inference (BART-MNLI) [17,18].

The TF-IDF model provides a transparent lexical lower bound: if it performs well, much of the task can be solved from surface words and phrases. BERT-base represents the supervised contextual approach used as the main model; if it substantially improves upon TF-IDF, then context-sensitive representation learning adds value beyond keyword matching. BART-MNLI represents a zero-shot large-model route in which labels are expressed as natural language hypotheses rather than learned from task-specific disaster labels. The zero-shot baseline is evaluated on the full 2879-post test split using the same six label descriptions as the supervised models.

For Transformer experiments, the random seed is fixed at 42 wherever stochastic training or sampling is used. BERT-base is fine-tuned for three epochs with a maximum sequence length of 128, batch size of 16, learning rate of 2×10^{-5}, and a fixed decision threshold of 0.5. The validation split is used for checkpoint selection based on Macro-F1; the held-out test split is used only for final reporting. The same 0.5 threshold is used for direct cross-lingual, multilingual, and translation-bridge prediction.

For the Chinese boundary test, a multilingual BERT model is additionally fine-tuned on the English main dataset and then transferred to Chinese social media text without any Chinese training data [9–12]. A translation-bridge baseline is also added: the Chinese posts are first translated into English and then passed to the same English BERT-base classifier used in the main experiment. Together, these comparisons clarify which part of the workflow is responsible for any performance loss: the classifier, the language mismatch, or the lack of task-specific Chinese training data.

3.5. Selective Adjudication Simulation

To examine human-in-the-loop updating under limited review resources, the training data are divided into a small seed set and a large unlabeled pool. The seed set contains 506 posts, corresponding to 5% of the original training portion, while the remaining 9568 posts form the adjudication pool. Four strategies are compared: random selection, uncertainty-based selection, disagreement-based selection, and selective adjudication. The purpose is not to simulate a complete annotation platform, but to test whether the same review budget produces different model improvement depending on how review candidates are selected. In this simulation, the labels of selected pool items are revealed from the already converted gold training data; no model-generated pseudo-labels are treated as human labels. The seed set is drawn from the training split with a 5% target ratio and a minimum of 20 examples per label, yielding 506 seed rows and 9568 pool rows. The committee consists of two one-vs-rest logistic-regression classifiers with balanced class weights: a word-level TF-IDF model using 1–2 g and a character-level TF-IDF model using character n-grams of length 3–5. The selective score uses the fixed weights 0.55 for uncertainty, 0.25 for disagreement, and 0.20 for their interaction. The event-group cap allows, at most, 80% of one 200-row batch to come from a single proxy event group.

Selective adjudication combines uncertainty and model disagreement so that human review is concentrated on posts that are both hard to classify and likely to change the decision boundary [13–15]. For a pool item (x_i), uncertainty is computed from the margin between the two largest normalized class probabilities:

$$u_i = 1 - \left(p_{i,(1)} - p_{i,(2)} \right). \quad (1)$$

Disagreement combines the average probability gap and the binary label disagreement between a word-level TF-IDF classifier and a character-level committee classifier:

$$d_i = 0.7\,\mathrm{mean}\left(|p_i^{word} - p_i^{char}|\right) + 0.3\,\mathrm{mean}\left(\hat{y}_i^{word} \neq \hat{y}_i^{char}\right). \quad (2)$$

After min–max scaling, the final selective score is:

$$s_i = 0.55\widetilde{u}_i + 0.25\widetilde{d}_i + 0.20\widetilde{u}_i\widetilde{d}_i. \quad (3)$$

The highest-scoring posts are selected for simulated human adjudication, with a simple proxy-event cap used to avoid domination of a review batch by one event group. The experiment proceeds in three rounds, with 200 adjudicated posts added per round, for a total human review budget of 600 posts. After each round, the classifier is retrained and evaluated on both the main test set and HumAID. This module therefore turns human review into an explicit, reproducible selection policy rather than an informal request for more labels.

3.6. Implementation and Reproducibility Details

The main BERT model uses bert-base-uncased; the multilingual comparison uses bert-base-multilingual-cased; the zero-shot NLI comparison uses a local bart-large-mnli pipeline. The main training script is scripts/train_bert.py, with data/processed/crisissense_converted.csv as the input and outputs/bert_event_group_devtest/best_model as the selected checkpoint. BERT and mBERT are trained for three epochs with a maximum sequence length 128, batch size 16, learning rate 2×10^{-5}, threshold 0.5, and random seed 42. The validation split is used for checkpoint selection by Macro-F1, and the test split is used only for final reporting. No class weighting is used in the Transformer fine-tuning runs; the selective-adjudication logistic-regression classifiers use balanced class weights. The optimizer is the default AdamW optimizer used by Hugging Face Trainer, with the learning rate specified above. The recorded environment is Python 3.13.9, PyTorch 2.6.0+cu124, CUDA 12.4, and one NVIDIA GeForce RTX 4060 Laptop GPU with 8 GB memory. All reported Transformer results are single-seed runs with seed 42 rather than mean $\pm$ standard deviation over repeated runs; this constraint is reflected in the limitations.

4. Results

4.1. Main Task Performance

Table 2 reports the main experiment on the held-out test split under the proxy event-group protocol. This is the central comparison of the paper because it answers the first research question: whether supervised contextual modeling remains necessary when the evaluation split is more conservative than a random text-level split. The validation split is not reused for final reporting, so the table reflects performance on unseen proxy event groups.

Table 2. Aggregate classification metrics on the held-out proxy event-group test split.

Method	Subset Acc	Macro-F1	Micro-F1
TF-IDF + Logistic Regression	0.6641	0.6133	0.7784
BERT-base	**0.9076**	**0.8824**	**0.9157**
Zero-shot NLI	0.0570	0.3581	0.3623

Note: Bold values indicate the best result for each metric.

BERT-base clearly outperforms both the sparse baseline and the zero-shot model on all reported metrics. Its Macro-F1 reaches 0.8824, compared with 0.6133 for TF-IDF plus logistic regression and 0.3581 for zero-shot inference. The gain over the sparse baseline is 0.2691 Macro-F1 points, indicating that contextual semantic modeling is critical for fine-grained disaster text discrimination. The gap between TF-IDF Micro-F1 (0.7784) and Macro-F1 (0.6133) also shows that lexical features handle frequent or easier categories better than balanced class-level performance. In contrast, BERT-base keeps both Micro-F1 and Macro-F1 high, suggesting that it improves not only the dominant categories but also the lower-frequency or semantically harder ones.

The zero-shot result is especially informative. Although zero-shot NLI produces a non-trivial Macro-F1 of 0.3581, its exact-match accuracy is only 0.0570. This means that the model sometimes assigns labels that overlap with part of the semantic space, but it does not reliably choose the correct operational category for individual posts. For emergency triage, this distinction matters because an apparently usable aggregate score can still conceal poor case-level routing.

Figure 2 visually confirms the ranking in Table 2 and makes the size of the performance gap clearer. BERT-base is separated from the two alternatives on every metric, while the zero-shot model remains far below the supervised methods. The figure therefore supports the interpretation that the advantage of supervised contextual modeling persists even under a stricter split protocol designed to reduce event overlap. This is important because it indicates that the main model is not simply benefiting from repeated lexical patterns within the same event cluster.

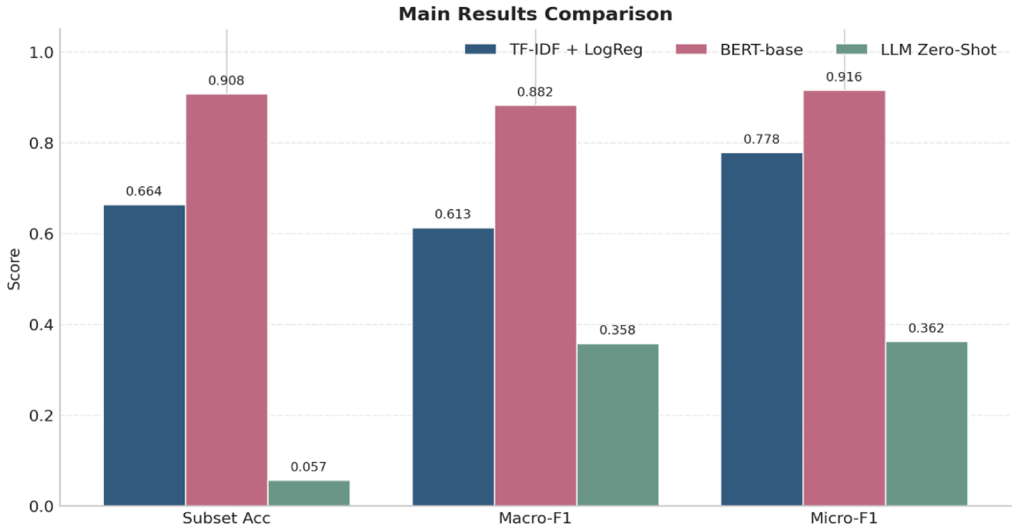

Figure 2. Main results under the proxy event-group split.

4.2. Per-Class Results and Error Profile

The per-class analysis addresses the second research question: which semantic categories remain difficult after fine-tuning. The per-class F1 values of BERT-base on the main test set are 0.9509 for casualty, 0.9125 for infrastructure damage, 0.7123 for request for help, 0.9351 for rescue or aid, 0.9094 for shelter or supply, and 0.8743 for situation update. The strongest performance is observed for categories with relatively explicit incident or response language, such as casualty and rescue or aid. The weakest class is request for help, which often appears through indirect, colloquial, or context-dependent phrasing. Figure 3 visualizes these class-level differences and helps identify which labels require the most caution in downstream use.

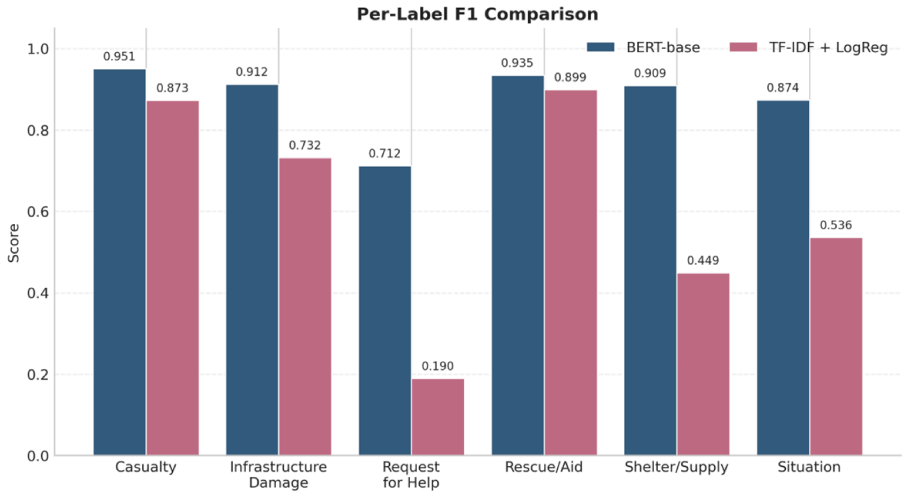

Figure 3. Per-class F1 comparison on the main task.

Figure 3 shows that the overall improvement in Table 2 is not driven by a single easy category. BERT-base improves over the sparse baseline across all six categories, with especially large gains for request for help (0.1898 to 0.7123), shelter or supply (0.4495 to 0.9094), and situation update (0.5359 to 0.8743). These are precisely the categories where surface keywords are often insufficient: requests may be implied rather than directly stated, shelter and supply needs may be expressed with varied local wording, and situation updates may overlap with several other labels. The largest remaining practical difficulty is therefore the boundary between implicit need expressions and general situation reports. This distinction is operationally relevant because emergency managers often care most about posts that imply unmet needs but do not state them in a rigid template.

4.3. Cross-Lingual Boundary Test

Table 3 begins the answer to the third research question by testing whether the trained model retains useful performance under Chinese boundary testing. It compares direct English-to-Chinese transfer, multilingual transfer, and a translation-bridge variant. This experiment is deliberately framed as a boundary test. The Chinese set is small, and no Chinese posts are used for training or tuning; therefore, the purpose is to test whether the English-trained semantic pipeline breaks under language shift, not to claim a deployable Chinese model.

Table 3. Cross-lingual boundary-test metrics for the Chinese social media set.

Model	Training Language	Chinese Subset Acc	Chinese Macro-F1	Chinese Micro-F1
English BERT-base	English	0.0960	0.0522	0.0965
Multilingual BERT	English	0.1921	0.2684	0.3632
Chinese-to-English translation + BERT-base	English after translation	**0.2090**	**0.3603**	**0.4784**

Note: Bold values indicate the best result for each metric.

The English-only BERT model is effectively unusable on the Chinese social media set, with a Macro-F1 of 0.0522 and a Micro-F1 of 0.0965. Its non-zero score is mainly attributable to broad situation-update predictions and does not indicate robust cross-lingual understanding.

By contrast, multilingual BERT reaches a Macro-F1 of 0.2684 and a Micro-F1 of 0.3632 on the same Chinese set. Translating the Chinese posts into English before applying the English BERT-base classifier gives the strongest result in this boundary test, with a Macro-F1

of 0.3603 and a Micro-F1 of 0.4784. This performance is still far below English in-domain performance, but it shows that translation can recover additional semantic alignment without adding Chinese posts to classifier training.

The comparison separates three sources of difficulty. First, the poor English-only result shows that the classifier cannot simply be applied to Chinese text without language support. Second, the mBERT result shows that multilingual pretraining partially restores semantic alignment, even though the model is fine-tuned only on English disaster data. Third, the translation bridge performs best overall, which suggests that converting the Chinese input into the language of the main classifier can recover useful semantic cues. However, even the best Chinese Macro-F1 (0.3603) remains far below the English main-test Macro-F1 (0.8824), so this should be interpreted as partial recovery rather than cross-lingual readiness.

Figure 4 makes the cross-lingual pattern visually clear: the direct English model is near failure, mBERT provides a moderate recovery, and translation bridging provides the strongest recovery among the three tested options. The result indicates that the main barrier is linguistic mismatch rather than a universal inability of Transformer models to represent disaster semantics. At the same time, the remaining gap between English and translated-Chinese performance shows that translation alone is not a full substitute for native Chinese training data or task-specific multilingual adaptation.

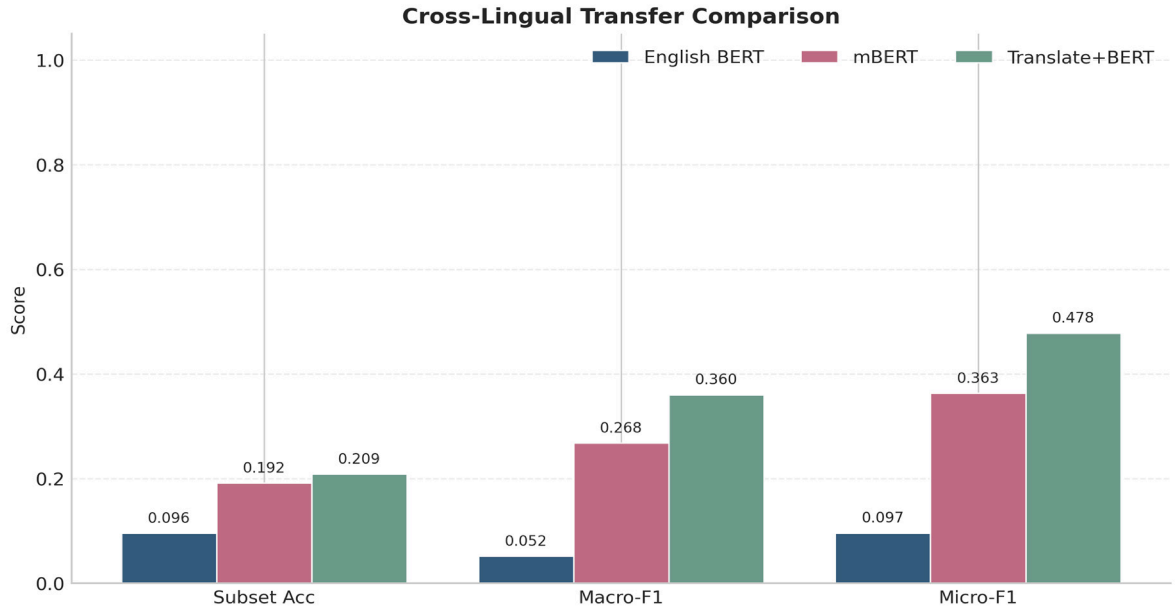

Figure 4. Cross-lingual transfer results on the Chinese social media set.

4.4. External Validation on HumAID

Table 4 completes the answer to the third research question by testing external English-dataset transfer. This experiment addresses a different form of generalization from the Chinese boundary test. Here, the language remains English, but the dataset source, event composition, and annotation conventions change. Without any retraining on HumAID, the model reaches a Macro-F1 of 0.8132 and a Micro-F1 of 0.8217. Relative to the main test performance, this corresponds to retaining more than 92% of the main-task Macro-F1, which supports the claim that the model learns transportable semantic signals rather than merely memorizing one dataset.

The drop from 0.8824 to 0.8132 Macro-F1 is expected because HumAID was not used during training and its labels must be aligned to the six-category schema. The important point is that the decrease is moderate rather than catastrophic. This result strengthens the

central argument of the paper: the six-category label design and supervised contextual classifier are not only fitting the converted CrisisSense-LLM data, but also capture a portion of disaster semantics that transfers to another established benchmark.

Table 4. External HumAID validation metrics for the main BERT-base model.

Dataset	Subset Acc	Macro-F1	Micro-F1
Main test split	0.9076	0.8824	0.9157
HumAID test split	0.8039	0.8132	0.8217

Figure 5 shows that the external validation loss is consistent across the aggregate metrics rather than being limited to one score. The largest category-level degradation appears in situation update, for which F1 decreases to 0.6540 on HumAID. This suggests that broader, more loosely bounded labels are more sensitive to differences in annotation style and event narration across corpora. Nevertheless, the model remains strong for casualty, rescue or aid, and shelter or supply, which are often the most practically actionable classes. In applied use, this pattern suggests that high-confidence outputs for concrete event and response categories are more dependable than broad situational summaries.

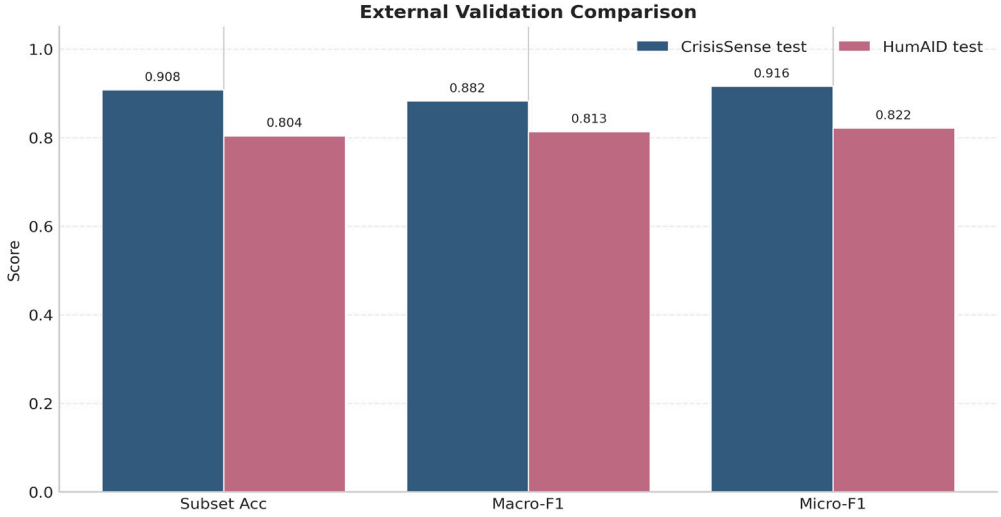

Figure 5. External validation results of the main model.

4.5. Selective Adjudication Under a Fixed Review Budget

Table 5 compares four adjudication strategies after three rounds of updating under the same review budget. The seed model starts at 0.7036 Macro-F1 on the main test split and 0.6369 on HumAID. After 600 adjudicated posts, selective adjudication produces the best final performance on both evaluation sets, reaching 0.7792 on the main test split and 0.7153 on HumAID. This result addresses the fourth research question: under a fixed human review budget, the ordering of reviewed samples matters.

Table 5. Results of budget-constrained selective adjudication.

Strategy	Initial Labeled Size	Added Reviews	Main-Test Macro-F1	HumAID Macro-F1
Random	506	600	0.7470	0.6941
Uncertainty	506	600	0.7781	0.7076
Disagreement	506	600	0.7681	0.7131
Selective adjudication	506	600	**0.7792**	**0.7153**

Note: Bold values indicate the best Macro-F1 in each evaluation setting.

The table also shows that the best strategy is not simply the one that chooses uncertain samples. Uncertainty sampling is nearly tied with selective adjudication on the main test set (0.7781 versus 0.7792), but selective adjudication is stronger on HumAID (0.7153 versus 0.7076). Disagreement sampling is weaker on the main test set but relatively competitive on HumAID. This pattern suggests that uncertainty and disagreement capture complementary information: uncertainty identifies samples near the current decision boundary, while disagreement can expose examples for which alternative feature views lead to different semantic decisions.

The strategies also differ in computational cost. Random selection requires no scoring pass. Uncertainty sampling requires probabilities from the word-level TF-IDF model. Disagreement and selective adjudication require probability passes from both the word-level and character-level committee models, followed by score sorting; selective adjudication adds only min–max scaling, an interaction term, and the event-group cap. All four strategies use the same human review budget of 600 posts and the same three retraining cycles. The reported advantage should therefore be interpreted as label efficiency under modest additional scoring cost, not as a claim of lower wall-clock time.

Figure 6 is useful because it shows the learning process, not only the final row of Table 5. The main-test gain of selective adjudication over the seed model is 0.0756, compared with 0.0434 for random adjudication. Expressed per 100 reviewed posts, the Macro-F1 gain is approximately 0.0126 for selective adjudication and 0.0072 for random adjudication. The advantage is therefore not only visible in the final score, but also meaningful in terms of review efficiency. In practice, this means that when annotation time is limited, reviewing the most informative posts can produce more improvement than simply increasing the amount of reviewed data.

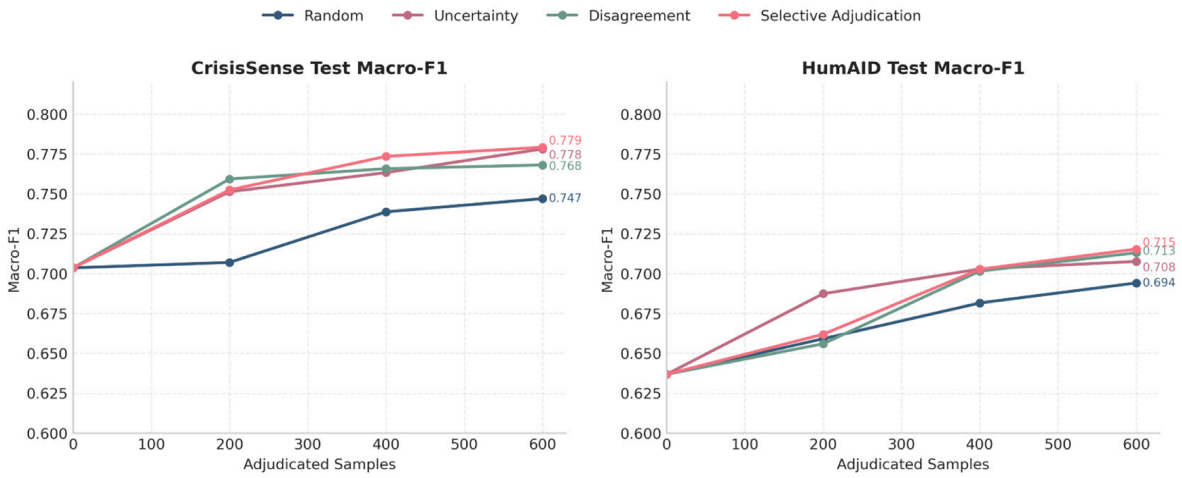

Figure 6. Performance curves for budget-constrained selective adjudication.

5. Discussion

5.1. Implications for Emergency Management

The main practical message of this study is that effective disaster text analysis does not necessarily require a novel architecture. What matters first is a task definition aligned with operationally meaningful categories and an evaluation protocol that avoids inflated performance estimates. The strong performance of BERT-base on the main task and its stable transfer to HumAID suggest that a well-trained supervised model can already support triage of actionable disaster content when the label space is carefully defined.

Figure 7 illustrates how the offline workflow could be embedded in an emergency-management setting. The classifier is not intended to issue autonomous operational com-

mands. Instead, it would prioritize high-volume social media streams into categories that can be inspected by analysts and routed to rescue, shelter, infrastructure-repair, public-warning, or information-verification teams. The selective review module then returns difficult cases to the labeled pool, creating an audit trail for future model updates.

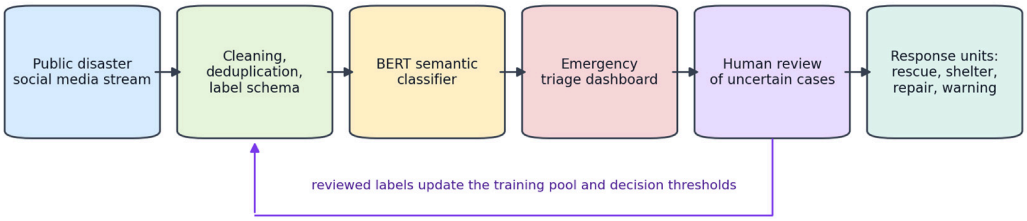

Figure 7. Emergency-response integration pathway for the proposed AI pipeline. Social media posts are cleaned, classified, routed to an emergency triage dashboard, reviewed by human analysts when uncertainty is high, and then connected to operational response units while reviewed labels update the data pool.

The class-level results are also informative for real-world use. Categories such as casualty, rescue or aid, and shelter or supply appear to be more stable and more portable across datasets, whereas requests for help and situation updates remain harder because their boundaries are broader and more context dependent. For emergency management, this means that model outputs should be interpreted asymmetrically: some categories may already be useful for decision support, while others still benefit from human confirmation.

The cross-lingual findings add a further deployment caution. A model that performs well on English disaster tweets should not be assumed to generalize to Chinese posts merely because the underlying model architecture is based on Transformers. The translation-bridge result is useful as an interim option, but the remaining performance gap shows that a Chinese-ready system would still require native Chinese data, careful label review, and likely stronger multilingual adaptation. Automatic translation can also accumulate errors before classification: local place names may be mistranslated, negations about casualties or damage may be weakened, and urgent help requests may lose pragmatic force. For this reason, translation bridging should be treated as a temporary diagnostic baseline rather than as a replacement for native Chinese annotation and model adaptation.

5.2. Methodological Implications

This study also highlights the importance of evaluation design. A model that performs well under a random split may still rely too heavily on repeated expressions from the same event. By contrast, the proxy event-group split used here makes the problem harder and the resulting claims more credible. Although the split is not a perfect event-level partition, it is a meaningful step toward more realistic cross-event testing.

The comparison with zero-shot inference further suggests that disaster-related short text remains a difficult setting for generic prompting-based approaches. In the present data conditions, supervised fine-tuning remains the more reliable option. In addition, the selective adjudication experiment shows that human review should not be treated as a simple quantity problem. The ordering of reviewed cases matters. When the same budget is spent on more informative cases, both in-domain and external performance improve more efficiently.

Taken together, these findings support a workflow view of disaster text classification. Model architecture is only one component. The credibility of the final system also depends on how labels are defined, how train-test leakage is controlled, whether external datasets are used for validation, whether language transfer is explicitly stress-tested, and whether human review is allocated to the most informative cases.

5.3. Limitations

The results should be interpreted within three boundaries. First, the study focuses on six relatively stable semantic categories, and the converted main corpus behaves mostly as a single-primary-label dataset; the findings therefore refer to fine-grained semantic discrimination rather than full multi-label dependency modeling. Second, the proxy event-group split reduces but does not fully eliminate ambiguity in event grouping. Third, the manually reviewed Chinese set is intentionally small and should be treated as a boundary test: it is sufficient to reveal the failure of direct English-to-Chinese transfer and partial recovery through multilingual or translation-bridged routes, but not to rank Chinese-ready systems with high confidence. Fourth, the Chinese set is manually reviewed but small, and no formal inter-annotator agreement was computed; it is therefore suitable for boundary testing but not for ranking deployable Chinese systems. Fifth, the Transformer experiments are single-seed runs with seed 42, so the robustness of small differences, especially in the selective-adjudication comparison, should be interpreted cautiously. Sixth, the multilingual comparison includes mBERT and a translation bridge but not stronger multilingual encoders such as XLM-R; testing such encoders is an important next step before any Chinese-ready deployment claim.

6. Conclusions

This study presented a complete evaluation pipeline for fine-grained semantic classification of disaster-related social media text for emergency management. Using a six-category task derived from CrisisSense-LLM, it showed that supervised BERT-base fine-tuning substantially outperforms both a sparse lexical baseline and zero-shot inference under a stricter proxy event-group split. The same model also transfers well to the mapped HumAID benchmark, indicating that the learned semantic representation is not confined to a single dataset.

The cross-lingual experiments showed that direct transfer from English to Chinese is ineffective with an English-only model, whereas multilingual pretraining and translation-bridged prediction provide limited but clear recovery. The human-in-the-loop experiments further showed, in a deterministic fixed-budget simulation, that selective adjudication can use the same review budget more efficiently than random sample addition and can improve both in-domain and external performance.

Overall, the evidence suggests that a practical disaster text analysis pipeline should combine supervised fine-grained classification, conservative evaluation design, external validation, and information-driven human review. Future work should expand the Chinese evaluation setting, test stronger multilingual encoders such as XLM-R, and move from offline adjudication simulation to fully operational annotation workflows. Additional per-class results and figure files are provided in the Supplementary Materials, and additional data and evaluation details are provided in Appendix A.

Supplementary Materials: The following supporting information can be downloaded at: https://github.com/HAUSTCourse/FGSC (accessed on 2 June 2026), Supplementary Materials include a per-class results table for the compared models and the figure files used in the manuscript. The code, validation data, exported tables, and figure files are available in the public project repository.

Author Contributions: Conceptualization, X.W. and T.Y.; methodology, X.W. and T.Y.; software, X.W.; validation, X.W.; formal analysis, X.W.; investigation, X.W.; resources, T.Y.; data curation, X.W.; writing—original draft preparation, X.W.; writing—review and editing, X.W. and T.Y.; visualization, X.W.; supervision, T.Y.; project administration, T.Y.; funding acquisition, T.Y. All authors have read and agreed to the published version of the manuscript.

Funding: The project was supported by the National Key R&D Program of China (No. 2025YFE0211300) and the National Earth Observation Data Center (No. NODAOP2025007).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The benchmark datasets analyzed in this study are publicly available from their original sources: CrisisSense-LLM and HumAID. The code, processed validation materials, exported result tables, and figure-generation files supporting the findings of this study are available in the project repository at https://github.com/HAUSTCourse/FGSC (accessed on 2 June 2026). The manually reviewed Chinese social media evaluation set is shared for research verification in a form consistent with platform terms of use and applicable data protection requirements. Raw third-party benchmark data remain governed by their original providers. The released repository avoids model weights, API keys, personal temporary files, and unnecessary user-identifying metadata; processed examples are provided only to the extent needed to verify the reported experiments.

Acknowledgments: The authors thank the maintainers of the public disaster datasets used in this study and the colleagues who assisted with the review of the Chinese evaluation sheet.

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

Appendix A. Additional Data and Evaluation Details

Table A1. Main CrisisSense-LLM split distribution by target category.

Split	Rows	Proxy Event Groups	Casualty	Infrastructure_ Damage	Request_for_ Help	Rescue_or_ Aid	Shelter_or_ Supply	Situation_ Update
Train	10,074	563	2073	1391	632	4492	536	950
Validation	1439	543	202	200	73	614	136	214
Test	2879	552	470	460	119	1184	303	343

Table A2. Chinese boundary-test composition.

Item	Value
Total posts	177
Informative posts	160
Non-informative/out-of-scope posts	17
Multi-label informative rows	105
Single-label informative rows	55
Weibo/Bilibili/Zhihu	145/23/9
casualty/infrastructure_damage/request_for_help	29/13/17
rescue_or_aid/shelter_or_supply/situation_update	130/69/65

Table A3. Selective-adjudication setup and relative computational cost.

Strategy	Scoring Information Used Before Each Review Batch	Additional Scoring Cost Before Selection	Human Reviews
Random	None	No scoring pass	600
Uncertainty	Word-level TF-IDF probabilities	One probability pass	600
Disagreement	Word-level and character-level TF-IDF probabilities	Two probability passes	600
Selective adjudication	Uncertainty, disagreement, interaction term, event-group cap	Two probability passes plus score scaling and sorting	600

References

1. Imran, M.; Castillo, C.; Diaz, F.; Vieweg, S. Processing Social Media Messages in Mass Emergency: A Survey. *ACM Comput. Surv.* **2015**, *47*, 67. [CrossRef] [PubMed]
2. Alam, F.; Qazi, U.; Imran, M.; Ofli, F. HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks. In *Proceedings of the International AAAI Conference on Web and Social Media*; AAAI Press: Washington, DC, USA, 2021; Volume 15, pp. 933–942. [CrossRef]
3. Yin, K.; Liu, C.; Mostafavi, A.; Hu, X. CrisisSense-LLM: Instruction Fine-Tuned Large Language Model for Multi-Label Social Media Text Classification in Disaster Informatics. *arXiv* **2024**, arXiv:2406.15477. [CrossRef]
4. Xie, S.; Hou, C.; Yu, H.; Zhang, Z.; Luo, X.; Zhu, N. Multi-Label Disaster Text Classification via Supervised Contrastive Learning for Social Media Data. *Comput. Electr. Eng.* **2022**, *104*, 108401. [CrossRef]
5. Olteanu, A.; Vieweg, S.; Castillo, C. What to Expect When the Unexpected Happens: Social Media Communications Across Crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*; ACM: Vancouver, BC, Canada, 2015; pp. 994–1009. [CrossRef]
6. Imran, M.; Mitra, P.; Srivastava, J. Enabling Rapid Classification of Social Media Communications During Crises. *Int. J. Inf. Syst. Crisis Response Manag.* **2016**, *8*, 17. [CrossRef]
7. Livers, J.; Johnson, N.; Spurlock, K.; Nasraoui, O. When Splits Matter: Interpreting Disaster Tweet Classification Models. In *Proceedings of the 2025 IEEE International Conference on Data Mining Workshops*; IEEE: New York, NY, USA, 2025. [CrossRef]
8. McDaniel, E.; Scheele, S.; Liu, J. Zero-Shot Classification of Crisis Tweets Using Instruction-Finetuned Large Language Models. In *Proceedings of the 2024 IEEE Global Humanitarian Technology Conference*; IEEE: New York, NY, USA, 2024. [CrossRef]
9. Pires, T.; Schlinger, E.; Garrette, D. How Multilingual Is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 4996–5001. [CrossRef]
10. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2020; pp. 8440–8451. [CrossRef]
11. Sánchez, C.; Sarmiento, H.; Abeliuk, A.; Pérez, J.; Poblete, B. Cross-Lingual and Cross-Domain Crisis Classification for Low-Resource Scenarios. In *Proceedings of the International AAAI Conference on Web and Social Media*; AAAI Press: Washington, DC, USA, 2023; Volume 17, pp. 754–765. [CrossRef]
12. Chowdhury, J.R.; Caragea, C.; Caragea, D. Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*; Association for Computational Linguistics: Florence, Italy, 2020; pp. 292–298. [CrossRef]
13. Lewis, D.D.; Gale, W.A. A Sequential Algorithm for Training Text Classifiers. In *SIGIR '94*; Springer: London, UK, 1994; pp. 3–12. [CrossRef]
14. Settles, B. *Active Learning*; Springer International Publishing: Cham, Switzerland, 2012. [CrossRef]
15. Freund, Y.; Seung, H.S.; Shamir, E.; Tishby, N. Selective Sampling Using the Query by Committee Algorithm. *Mach. Learn.* **1997**, *28*, 133–168. [CrossRef]
16. Olteanu, A.; Castillo, C.; Diaz, F.; Vieweg, S. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Proceedings of the International AAAI Conference on Web and Social Media*; AAAI Press: Washington, DC, USA, 2014; Volume 8, pp. 376–385. [CrossRef]
17. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Florence, Italy, 2020; pp. 7871–7880. [CrossRef]
18. Yin, W.; Hay, J.; Roth, D. Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 3912–3921. [CrossRef]
19. Pohl, D.; Bouchachia, A.; Hellwagner, H. Batch-Based Active Learning: Application to Social Media Data for Crisis Management. *Expert Syst. Appl.* **2018**, *93*, 232–244. [CrossRef]
20. Kaufhold, M.; Bayer, M.; Reuter, C. Rapid Relevance Classification of Social Media Posts in Disasters and Emergencies: A System and Evaluation Featuring Active, Incremental and Online Learning. *Inf. Process. Manag.* **2020**, *57*, 102132. [CrossRef]

21. Pandey, R.; Purohit, H.; Castillo, C.; Shalin, V.L. Modeling and Mitigating Human Annotation Errors to Design Efficient Stream Processing Systems with Human-in-the-Loop Machine Learning. *Int. J. Hum.-Comput. Stud.* **2022**, *161*, 102772. [CrossRef]
22. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; Volume 1 (Long and Short Papers), pp. 4171–4186. [CrossRef]