

Article

Zero-Shot Social Media Crisis Classification: A Training-Free Multimodal Approach

Franziska Schwarz ^{1,2,*} , Klaus Dieter Schwarz ² , Daniel Arias Aranda ¹ , Kendrick Bollens ² ,
Navaneeth Shivananjappa ³ , Reiner Creutzburg ^{3,4,5}  and Vesna Dimitrova ⁶ 

¹ Department of Business and Economics, University of Granada, 18071 Granada, Spain

² School of Technology and Architecture, SRH University of Applied Sciences Heidelberg, D-04317 Leipzig, Germany; kendrick.bollens@srh.de (K.B.)

³ School of Technology and Architecture, SRH University of Applied Sciences Heidelberg, D-12059 Berlin, Germany; creutzburg@th-brandenburg.de (R.C.)

⁴ Fachbereich Informatik und Medien, Technische Hochschule Brandenburg, D-14770 Brandenburg, Germany

⁵ Research Group “Applied Computer Science”, German University of Digital Science, D-14482 Potsdam, Germany

⁶ Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia; vesna.dimitrova@finki.ukim.mk

* Correspondence: fschwarz@correo.ugr.es

Abstract

Rapid classification of social media content during humanitarian crises is essential for effective disaster relief; however, traditional approaches require extensive annotated training data, which are often unavailable during new disasters. This paper presents a training-free, multimodal classification framework that leverages zero-shot vision-language models to analyze disaster-related social media content without task-specific training. The framework employs a two-stage prompt-engineered pipeline using the locally deployable Mistral-Small-3.1-24B-Instruct model, performing binary informativeness detection followed by multiclass categorization into eight humanitarian categories through structured JSON output generation. Evaluation on the CrisisMMD dataset of 18,082 multimodal samples from seven natural disasters demonstrated binary F1 scores above 0.84 for both text and image informativeness detection and weighted F1 scores of 0.61 (text) and 0.72 (image) for humanitarian categorization. The framework generalizes consistently across all disaster types with minimal performance variance (standard deviation below 0.031) and operates entirely on local infrastructure without cloud dependencies, requiring only moderate GPU resources. By eliminating training data requirements, this approach enables immediate deployment during new disasters, demonstrating that zero-shot multimodal classification achieves practically relevant performance for real-time crisis response.

Keywords: zero-shot classification; multimodal analysis; disaster management; vision-language models; social media; humanitarian crisis; emergency response; training-free classification



Academic Editor: Douglas O'Shaughnessy

Received: 1 February 2026

Revised: 15 February 2026

Accepted: 23 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The frequency and intensity of natural disasters are increasing at an alarming rate, worldwide. According to data from the United Nations Office for Disaster Risk Reduction (UNDRR), the number of global disasters could rise from around 400 per year in 2015 to 560 per year by 2030, equivalent to 1.5% of major disasters every day [1]. In the United States alone, the number of billion-dollar disasters rose from an average of 9.0 events per year (1980–2023) to 28 events in 2023, representing an increase of 156% [2]. This escalation

is exacerbated by climate change, with 2024 considered the hottest year on record at 1.55 °C above pre-industrial levels [3].

During these crises, social media established itself as a critical source of real-time information. People use platforms such as Twitter, Facebook, and WhatsApp to send calls for help, document damages, and share vital information [4,5]. The 2023 earthquake in Turkey and Syria demonstrated the life-saving importance of these communication channels when people trapped under rubble use social media to alert emergency responders [6]. Studies show that social media often reports local conditions faster and more accurately than traditional news sources [7].

However, the sheer volume of multimodal content generated far exceeds human processing capacity. Major disasters generate millions of posts, with Hurricane Sandy alone producing 52.5 million tweets from 13.75 million users [8]. Hurricane Ian in 2022 resulted in over 20 million analyzed tweets [9]. To make matters worse, only 8–20% of this content actually contains actionable information for disaster management; the rest consists of personal reactions, general information exchange, and unusable content [10,11]. This challenge is exacerbated by the multimodal nature of the data: users share text messages, as well as images and videos that may contain critical visual information about infrastructure damage, injured people, or blocked rescue routes [12].

Traditional machine learning approaches have enabled important advances in disaster analysis, but they have reached practical limits in time-critical situations. These systems require extensive annotated training data, which is time-consuming and resource-intensive to create [8]. For example, annotating 25,000 disaster images requires more than 300 h of work, while expert assessment of building damage requires 40–80 h per 1000 images [13]. This requirement is particularly challenging for new or rare types of disasters for which sufficient historical data may not be available. Moreover, while first responders require actionable information within 30 min to 3 h, traditional ML pipelines typically require 12–44 h for deployment [2,14], with deep learning models for disaster image classification alone requiring 8–24 h of training [15]. These delays create a critical gap in the first 72 h when most lives can be saved.

Another aspect concerns the specialization of models for specific disaster types and geographic regions. A system trained on hurricane data may be less effective in analyzing earthquake damage because the visual and textual patterns between different disaster types vary significantly [16,17]. This domain specificity often requires the development of separate models for different types of disasters, which increases the maintenance costs.

Finally, trained models require continuous maintenance as linguistic changes, new platforms, and evolving user behavior degrade performance over time [18,19]. These challenges have motivated the search for more flexible and rapidly deployable approaches.

Zero-shot learning approaches offer promising opportunities to address time-critical challenges in disaster analysis. The key advantage is their immediate deployability, without task-specific training [20]. Whereas traditional approaches require weeks to months for data collection, annotation, and model training, zero-shot systems can be deployed within minutes to hours [21]. This speed is particularly valuable in the critical first hours of a disaster, when quick decisions can save lives.

Another significant advantage is the ability to generalize to new and unforeseen types of disasters. Zero-shot models can classify events for which they have never been explicitly trained by drawing on the broad knowledge acquired in the pre-training phase [22,23]. This enables immediate application to novel disasters, whether a rare volcanic eruption, an unexpected industrial accident, or disasters in geographic regions for which no training data exist [24]. Studies have shown that zero-shot approaches can achieve consistent performance across different types of disasters, from hurricanes to earthquakes and floods [25].

Recent advances in vision-language models have made these capabilities possible. Modern models such as CLIP, ALIGN, and newer architectures have been pre-trained on large numbers of image-text pairs, developing a deep understanding of the relationships between visual and linguistic concepts [26,27]. These models can directly link natural language descriptions of disaster scenarios to visual patterns in images, without requiring explicit training examples [28]. Current benchmarks show impressive results: zero-shot approaches achieve up to 98.83% accuracy in emergency vehicle detection [29] and 69–86% in zero-shot transfer damage assessment [30] without task-specific training on these domains.

This study extends the Open-Source Intelligence Disaster Event Tracker (ODET) system [31] with multimodal analysis capabilities through zero-shot classification. ODET was originally developed as a modular platform for automated disaster reporting from social media text data, and has already used zero-shot classification for text analysis. This study demonstrates how this architecture can be successfully extended to include image analysis capabilities without losing the fundamental advantages of the training-free approach.

Our experimental results on the CrisisMMD dataset demonstrate that zero-shot classification achieves practically relevant performance values. For informativeness detection, binary F1 scores of 0.8415 and 0.8417 were obtained for text and image content, respectively. The nearly identical values across both modalities demonstrate the robustness of this approach. For the more challenging task of humanitarian categorization into eight classes, weighted F1 scores were achieved of 0.6106 (text) and 0.7213 (image), with better performance on images owing to their often more explicit visual information.

Validation across seven different natural disasters, including hurricanes Harvey, Irma, and Maria, earthquakes in Mexico and Iraq-Iran, wildfires in California, and floods in Sri Lanka, show remarkably consistent results. The standard deviation of the F1 scores was only 0.031 for text and 0.024 for images, indicating the strong generalization ability of the approach. It is particularly noteworthy that smaller datasets, such as the Iraq-Iran earthquake (498 samples), achieve comparable or even better results than larger datasets, demonstrating independence from the dataset size.

A key practical contribution of this study is the demonstration that such a system can be implemented with moderate effort. The development of the multimodal extension took approximately two weeks, with the main efforts being prompt optimization, JSON schema definition, and API integration. Importantly, no special machine learning expertise or complex training procedures are required. The implementation uses publicly available models (Mistral-Small-3.1-24B-Instruct) and standard libraries to ensure reproducibility and accessibility for organizations with limited technical resources.

This study addressed three central research questions regarding the practicality of zero-shot multimodal classification in disaster management.

1. **Can zero-shot classification achieve practically relevant performance for the analysis of disaster content?** This question investigates whether training-free approaches can offer sufficient accuracy for real-world use in emergencies. Although academic benchmarks are important, the performance values achieved must also meet the practical requirements of first responders and disaster managers. This is evaluated using established metrics (F1 score, precision, recall) for both classification tasks: binary informativeness detection and multi-class categorization of humanitarian content.
2. **How well does the approach generalize across different types of disasters?** The ability to generalize is crucial for practical utility as each disaster has unique characteristics. A system that works only for certain types of disaster would be of limited value in the dynamic reality of emergency management. The consistency of model performance is investigated across seven different natural disasters with varying visual and textual characteristics, from rapidly developing hurricanes to sudden earthquakes.

3. **What is the actual implementation effort required?** In addition to technical performance, practical feasibility is crucial. Many promising research approaches fail because of the complexity of their implementation. This question examines the actual time and resource requirements, technical skills required, and practical challenges of integrating zero-shot systems into existing disaster management workflows. The answer to this question determines whether the approach is accessible to organizations with limited technical resources.

Based on these research questions, this study makes three principal contributions. First, it demonstrates that zero-shot multimodal classification achieves practically relevant performance for disaster content analysis, with binary F1 scores above 0.84 for both text and image informativeness detection and weighted F1 scores of 0.61 (text) and 0.72 (image) for humanitarian categorization, without any task-specific training. Second, it validates consistent cross-disaster generalization across seven distinct natural disasters with minimal performance variance (standard deviation below 0.031), confirming that the framework can be deployed to new, unforeseen disaster types without adaptation. Third, it provides empirical evidence that the primary implementation effort lies in prompt engineering and system integration rather than machine learning expertise, requiring approximately two weeks of development using publicly available models deployable on local infrastructure with moderate GPU resources.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive overview of related studies in three areas: social media analysis in disaster management, zero-shot multimodal classification, and the use of multimodal language models in disaster relief. Section 3 describes the materials and methods used, including the CrisisMMD dataset with 18,082 multimodal samples, configuration of the Mistral-Small-3.1-24B-Instruct model, and the two-stage classification methodology. Section 4 presents the results of the experiments, which demonstrate strong performance and consistent generalization across different disaster types. Section 5 summarizes the findings, discusses the practical implications for disaster management, and identifies promising directions for future research.

2. Related Work

Our research builds on earlier work on social media analysis in disaster response, zero-shot multimodal classification, and multimodal-language models in disaster management.

2.1. Social Media Analysis in Disaster Response

Social media analysis in disaster response enables rapid situational awareness and damage assessment, allowing authorities to respond to emergencies [4]. This research addresses the challenge of filtering irrelevant images and estimating disaster severity from social media content, which is typically high in volume and velocity [1–5,7,8,10–12,32–34]. By analyzing both text and image modalities, these studies aim to extract valuable information from social media data to support disaster response efforts.

Gupta et al. [4] developed a framework to filter irrelevant images and assess disaster impact on social media, aiming to reduce computational cost while maintaining comparable accuracy to deep learning models. Ponce-López et al. [5] employed binary classification methods, including DistilBERT, to evaluate disaster severity and precision in social media data. Imran et al. [8] utilized convolutional neural networks to create a multimodal deep learning architecture for analyzing social media data during disasters. Alharbi et al. [2] demonstrated the practical advantages of DistilBERT in emergency management by achieving significant predictive accuracy through iterative model refinement. Karimiziarani [1] reviewed the current state of social media analytics in disaster management, highlighting

the need for continued research and innovation. Temkar et al. [10] focused on real-time data collection and analysis from social media platforms to identify disaster types, estimate severity, and provide actionable insights. Firmansyah et al. [7] explored automatically extracting textual information from social media images to enhance disaster area mapping. He et al. [3] presented an AI-powered framework combining natural language processing and geospatial visualization to analyze disaster-related social media content. Ponce-López et al. [32] aligned behavioral indicators with climate variables to provide additional information during different phases of disasters. Hanny et al. [11] introduced a novel approach for multimodal information extraction from geo-social media data to aid decision support in disaster management. Beyond disaster-specific analysis, research has also demonstrated the broader utility of social media as real-time sensors for crisis-related events. Aljurbua et al. [33] developed a spatio-temporal multiplex network leveraging social media data for weather-caused power outage prediction, while Otudi et al. [34] applied social sensor data and social network analysis to classify severe weather events.

To summarize, recent research has focused on extracting valuable information from social media data to support disaster response efforts, with studies analyzing both text and image modalities. Key approaches include image filtering and impact assessment [4], binary classification methods for severity evaluation [5], multimodal deep learning architectures [8], and AI-powered frameworks for real-time analysis and geospatial visualization [3]. Additionally, research has explored automatic textual information extraction from social media images [7] and the integration of behavioral indicators with climate variables [32].

2.2. Zero-Shot Multimodal Classification

Zero-shot multimodal classification enables the classification of data without task-specific training, allowing models to handle new tasks or datasets immediately upon deployment. This approach combines information from multiple modalities (such as text and images) to make predictions on classes that were not present during training. The technique has become particularly relevant in scenarios where labeled training data is scarce or unavailable, such as real-time crisis response, emerging disaster situations, and specialized medical applications where obtaining annotated samples is challenging.

Liu et al. [20] developed MERL, a multimodal model that enables zero-shot ECG classification using text prompts, eliminating the need for task-specific training data in medical diagnostics where annotated samples are often limited. Abdelhamed et al. [21] introduced an approach for zero-shot image classification using multimodal large language models (LLMs), achieving state-of-the-art accuracy on multiple computer vision datasets. Wang et al. [22] presented CapS-Adapter, a method that leverages caption-based support sets for training-free multimodal classification, demonstrating its effectiveness across 19 benchmark vision-language datasets. Zou et al. [35] proposed VERO, an acquisition function designed to select challenging samples for few-shot learning of large vision-language models (LVLMs). Hickmon [28] aimed to create a meta-model combining predictions of multiple open-vocabulary multimodal models without additional training or fine-tuning. Islam et al. [25] explored the efficacy of large multimodal models (LMMs) in performing image classification tasks using tailored prompts. Chen et al. [24] transformed high-resolution remote sensing image scene classification into a question-answering task using multimodal data, employing the open-source large multimodal model Llava. Udandara et al. [23] analyzed the extent to which pretraining datasets encompass downstream concepts targeted for zero-shot evaluation, addressing challenges in achieving zero-shot generalization under large-scale training paradigms.

To summarize, recent research has advanced zero-shot multimodal classification across diverse domains including medical diagnostics [20], computer vision [21], vision-language

tasks [22], and remote sensing [24]. Key developments include models that leverage text prompts for classification without training data, methods using caption-based support sets, and meta-models combining multiple predictions. While these approaches demonstrate the potential for immediate deployment on new tasks, challenges remain in understanding how pretraining datasets relate to downstream zero-shot capabilities [23].

2.3. Multimodal-Language Models in Disaster Management

Multimodal-language models in disaster management leverage textual and visual data from social media to provide rapid, immersive insights for emergency response, addressing the challenge of overwhelming irrelevant or redundant information [27]. These models aim to enhance disaster management by offering a comprehensive understanding of on-the-ground conditions and public responses, aiding in timely, informed decision-making [11]. By integrating textual, visual, and audio modalities, multimodal large language models (MLLMs) can automate safety-critical event detection and analysis, improving the reliability and accuracy of insights [36]. Furthermore, MLLMs can streamline disaster response activities, reducing human execution time and optimizing resource distribution, particularly in underdeveloped regions [13].

Hughes et al. [27] employed Gemini 1.5 to automate video identification during hurricanes on X (formerly Twitter). Lei et al. [14] surveyed and taxonomized LLMs in natural disaster management to guide professional community development. Kadiyala et al. [15] demonstrated GPT-4 Vision's proficiency in flood and water quality assessment, and MLLMs' potential in hydrological applications. Hanny et al. [11] introduced a clustering approach for extracting insights from geo-social media data in disaster management. Abu Tami et al. [36] presented the first MLLM for handling multiple tasks in safety-critical event management. Chen et al. [13] developed DisasTeller, a multi-LVLM-powered framework for automating post-disaster management tasks. Basit et al. [6] developed a hierarchical classification framework for disaster tweets achieving 91.53% accuracy for informativeness detection. Goecks and Waytowich [37] introduced DisasterResponseGPT leveraging in-context learning for rapid emergency plan generation. Kumbam et al. [38] presented FloodLense, a framework combining ChatGPT with computer vision models for real-time flood detection. Kalluri et al. [39] addressed data scarcity through synthetic data generation for disaster assessment from aerial imagery. Gao et al. [9] demonstrated real-time monitoring capabilities using ERNIE NLP models combined with Deeplab v3+ during Typhoon Haikui. Rawat [40] developed DisasterQA as a benchmark for assessing LLM performance in disaster response scenarios. Morocco Solidarity Hackathon [41] analyzed AI deployment during the Morocco earthquake response. Zou et al. [19] explored multimodal social media data fusion for disaster image classification. Zhang et al. [18] provided a comprehensive review of multi-source multimodal data and deep learning for disaster response. Imran et al. [8] examined opportunities and challenges in using AI and social media multimodal content for disaster management. Li et al. [26] demonstrated BLIP-2's efficiency improvements in vision-language models. An et al. [17] developed DAVI for generalizable disaster damage assessment using vision foundation models.

To summarize, multimodal language models are increasingly employed in disaster management to integrate textual, visual, and audio data from social media, enhancing situational awareness, automating event detection, and streamlining response activities. Key approaches include video identification during hurricanes [27], hierarchical classification frameworks for disaster tweets [6], real-time flood detection systems [38], and synthetic data generation for aerial imagery assessment [39]. Recent deployments demonstrate practical applications in the Morocco earthquake [41] and Typhoon Haikui [9], while benchmarks like DisasterQA [40] evaluate model performance in emergency scenarios. However, chal-

Challenges remain in computational requirements, data quality in crisis conditions, and ethical considerations for deployment in underdeveloped regions [8,13,18].

3. Materials and Methods

This chapter describes the materials and methods employed in developing and evaluating the zero-shot multimodal classification framework for content related to humanitarian crises. First, the CrisisMMD dataset, comprising 18,082 multimodal samples from seven natural disasters, was introduced as a benchmark for evaluation [42]. Next, the configuration of the Mistral-Small-3.1-24B-Instruct vision language model without task-specific training is explained. The methodology involves a two-stage classification approach that determines the informativeness and humanitarian categories of the text and image content. Particular emphasis was placed on prompt engineering and the use of structured JSON schemas to achieve consistent classification results across modalities. Finally, an evaluation methodology is presented that uses F1 scores, accuracy metrics, and confusion matrices to comprehensively assess performance.

3.1. Dataset Description

To evaluate the framework, the CrisisMMD (Crisis Multimodal Dataset) developed by the Qatar Computing Research Institute (QCRI) [42] was used. This dataset was specifically designed for multimodal analysis of disaster content in social media and represents an established benchmark in disaster management research [8]. The CrisisMMD dataset was selected because of its comprehensive annotations for both text and image modalities, which allowed for a direct evaluation of the multimodal zero-shot approach.

The dataset comprised 18,082 tweet-image pairs collected during seven major natural disasters in 2017. The data were collected on an event-specific basis using defined keywords and time periods: Hurricane Harvey (25 August to 20 September), Hurricane Irma (6–21 September), Hurricane Maria (20 September–13 November), Mexico Earthquake (20 September–6 October), California Wildfires (10–27 October), Iraq-Iran Earthquake (13–19 November), and Sri Lanka Floods (31 May–3 July). This systematic collection ensured the representative coverage of the respective disaster events throughout the course.

The annotation scheme for the CrisisMMD dataset was based on two hierarchical classification levels. The first level comprises binary informativeness labels that indicate whether a tweet or image contains relevant information about a disaster. Of the 16,058 annotated texts, 11,509 (71.7%) were classified as informative and 4549 (28.3%) as noninformative. The 18,082 images showed a more balanced ratio, with 9374 (51.8%) informative and 8708 (48.2%) non-informative content.

The second level categorizes informative content into eight humanitarian classes. The distribution shows significant class imbalances: in texts, “other relevant information” (5954 samples, 37.1%) and “rescue volunteering or donation effort” (3293 samples, 20.5%) dominate, while “missing or found people” (40 samples, 0.2%) and “vehicle damage” (54 samples, 0.3%) were severely underrepresented. The image annotations show similar imbalances, with “infrastructure and utility damage” (3624 samples, 20.0%) being the most frequent informative category, followed by “other relevant information” (2529 samples, 14.0%).

A key feature of the dataset is dual annotation, in which text and images are classified independently. This enables a differentiated analysis of modality-specific performance, and considers that text and images may contain different or complementary information. In addition, the dataset contains annotations on damage severity for 3526 images, divided into “little or no damage” (475 samples), “mild damage” (839 samples), and “severe

damage” (2212 samples). These annotations enable a more detailed analysis of the visual damage assessment but were not considered in the present study.

The pronounced class imbalance poses a particular challenge for zero-shot approaches because there is no possibility of training-based balancing. Therefore, the framework must correctly identify both common and rare categories without prior training.

It should be noted that the CrisisMMD dataset was collected during disasters in 2017, and social media behavior, platform features, and content formats have evolved since then. However, CrisisMMD remains the most widely used multimodal benchmark in crisis informatics research [8,43], and its use enables direct comparison with the extensive body of published results. The zero-shot nature of our approach, which relies on general language and vision understanding rather than platform-specific patterns, partially mitigates concerns about temporal generalizability. Nevertheless, evaluation on more recent disaster datasets should be pursued in future work.

3.2. Model Architecture and Configuration

For the zero-shot classification framework, the Mistral-Small-3.1-24B-Instruct model was used, an instruction-optimized large language model with native vision-language capabilities. The selection of this model was based on several key factors: first, it offers robust zero-shot classification capabilities without task-specific training; second, it supports structured output formats through JSON schema enforcement; and third, it enables seamless processing of multimodal inputs through integrated image processing capabilities.

The Mistral-Small-3.1-24B-Instruct model excels in following complex instructions accurately, while generating consistent, structured outputs. This property is particularly important for the application because classification must be performed into predefined categories without training examples. The instruction-tuning phase of the model was prepared to translate the natural language task descriptions directly into correct classification decisions.

For practical use, a 6-bit quantized version of the model was used (Q6_K_L), which offers an optimal compromise between the model quality and resource efficiency. This quantization reduces the model size to approximately 20 GB, with total runtime memory requirements of 25–27 GB including the multimodal projector and processing buffers, without causing any significant loss in classification accuracy. Quantization enables the deployment of consumer hardware with moderate GPU resources, thereby increasing the practical applicability of the framework.

The model configuration was specifically optimized for deterministic and consistent classification results. The context window was set to 24,576 tokens, which provided sufficient capacity for processing tweet text, system prompts, and structured output instructions. The temperature was set to zero to ensure maximum prediction consistency and eliminate random variations in the classification decisions. In addition, JSON schema validation ensures that all model outputs conform to the predefined structure format.

A key feature of the proposed approach is complete avoidance of training or fine-tuning. The model was used exclusively in its pre-trained form, with the classification task defined by carefully crafted prompts. This pure zero-shot approach demonstrates the ability of modern vision-language models to handle complex domain-specific tasks without additional training. The absence of a training process enables immediate application to new disaster events and eliminates the need for annotated training data, which is particularly advantageous in time-critical crises.

3.3. Classification Methodology

Our classification methodology is based on a modular pipeline architecture that enables systematic processing of multimodal disaster content. The pipeline comprises four main components: (1) data preprocessing, (2) prompt-based classification, (3) structured output generation, and (4) result validation. This architecture ensures consistent processing of both text and image content, while allowing flexibility for future extensions. Figure 1.

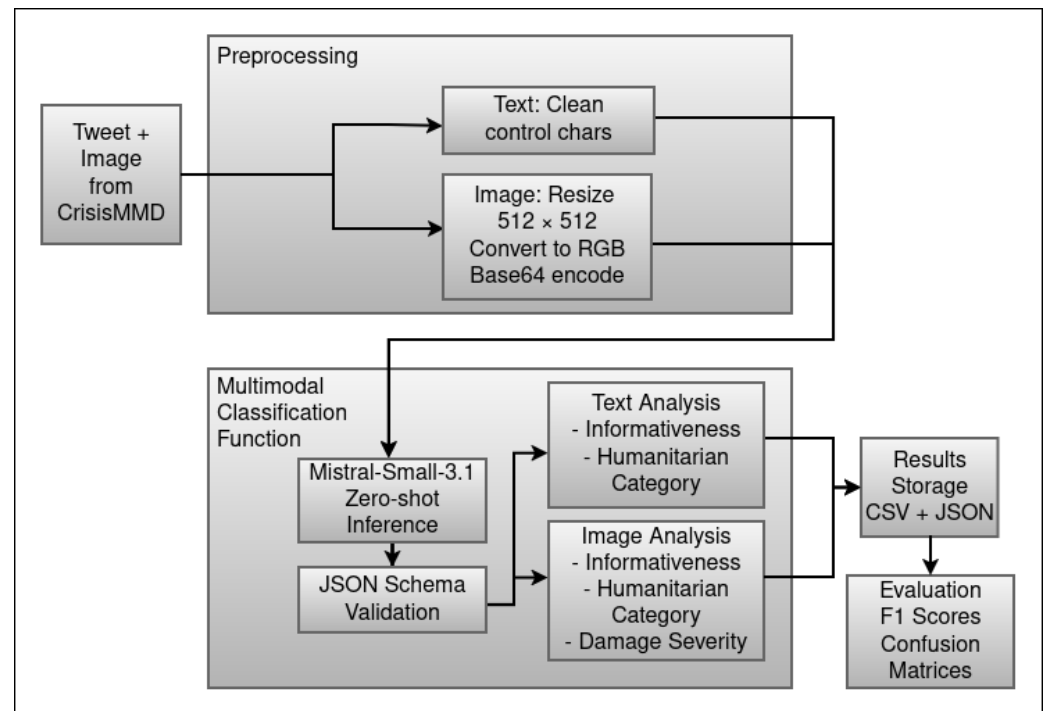


Figure 1. Preprocessing and Multimodal Classification.

The heart of the methodology is a prompt engineering approach, which is based on three fundamental design principles. First, explicit task-specific instructions were used to provide the model with clear classification criteria. Second, the prompts were structured hierarchically, first determining the informativeness and then the humanitarian category. Third, contextual information about disaster events was integrated to increase classification accuracy. These principles enabled the model to make consistent decisions without prior training.

We developed separate specialized prompts for the two-stage classification. Informativeness prompts the model to evaluate whether the content contains relevant disaster information, considering explicit criteria such as damage reports, offers of assistance, or situation reports as listed in Listing 1. The categorization prompted the model to assign informative content to one of the eight humanitarian categories by providing detailed descriptions and examples for each category. Both prompts were iteratively optimized to minimize ambiguity and maximize classification consistency.

Listing 1. Systematic classification instructions (excerpt).

```

1 CLASSIFICATION CRITERIA:
2 STEP 1: Analyze the TWEET TEXT first:
3 1. Tweet informativeness - determine if the tweet contains
   SPECIFIC information:
4     - informative: Contains SPECIFIC disaster impact/response
       evidence
5     - not_informative: Generic statements, emotions only,
       unrelated content
6
7 2. Tweet category - identify the DOMINANT content:
8     - affected_individuals: Mentions displaced people,
       survivors
9     - infrastructure_and_utility_damage: References damaged
       buildings
10    [... complete category definitions ...]

```

Cross-modality consistency was ensured through a uniform classification and category definitions. While the prompt formulations are adapted to the respective modality—for example, “tweet text” versus “image shows”—the underlying classification rules remain identical. This consistency enables direct comparisons between text and image classification and ensures that both modalities are evaluated according to the same humanitarian criteria.

Structured output generation is achieved through predefined JSON schemas that specify the exact structure of model outputs. The schema defines two main objects: “text_analysis” and “image_analysis”, each with the fields “informativeness” and “category”. For the image analysis, the field “damage_severity” was also included as listed in Listing 2. This schema-based output ensures parsable and validatable results, which significantly simplifies further automated processing.

The text-processing pipeline includes minimal pre-processing steps to preserve the original information. Tweet texts were passed directly to the model with only the control characters removed. This deliberate decision against extensive preprocessing was based on the observation that modern language models are robust to informal writing styles, hashtags, and emojis, which are common in disaster tweets.

The image-processing pipeline requires additional steps to prepare visual data. Images were first scaled to a uniform size of 512×512 pixels to ensure consistent inputs and optimize the memory requirements. The RGBA images were converted to RGB using a white background for transparency. This is followed by Base64 encoding in the JPEG format with a quality of 85%, which represents an optimal compromise between image quality and data size. The encoded images are integrated into multimodal prompts as data URLs.

Batch processing has been implemented to increase efficiency when analyzing large datasets. With a batch size of 16 samples, text-image pairs are processed in parallel, with error-handling mechanisms ensuring that individual processing errors do not affect the entire batch. The implementation uses ThreadPoolExecutor for parallel processing with 16 worker threads, enabling the optimal utilization of modern multicore systems. Intermediate results were periodically saved to prevent data loss during longer processing runs.

Listing 2. Complete JSON schema for bidirectional classification.

```

1 {
2   "type": "object",
3   "properties": {
4     "text_analysis": {
5       "type": "object",
6       "properties": {
7         "informativeness": {
8           "type": "string",
9           "enum": ["informative", "not_informative"]
10        },
11        "category": {
12          "type": "string",
13          "enum": ["affected_individuals", "
14                infrastructure_and_utility_damage",
15                "injured_or_dead_people", "
16                missing_or_found_people",
17                "not_humanitarian", "
18                other_relevant_information",
19                "rescue_volunteering_or_donation_effort", "
20                vehicle_damage"]
21        }
22      },
23      "required": ["informativeness", "category"]
24    },
25    "image_analysis": {
26      "type": "object",
27      "properties": {
28        "informativeness": {"type": "string", "enum": ["
29              informative", "not_informative"]},
30        "category": {"type": "string", "enum": [...]},
31        "damage_severity": {"type": "string",
32              "enum": ["little_or_no_damage", "
33              mild_damage",
34              "severe_damage", "
35              dont_know_or_cant_judge"
36              ]}
37      },
38      "required": ["informativeness", "category", "
39            damage_severity"]
40    }
41  },
42  "required": ["text_analysis", "image_analysis"]
43 }

```

3.4. Evaluation Framework

Our evaluation framework was developed to comprehensively assess the performance of a zero-shot classification system. The binary F1 score was employed as the primary metric, as it represents the harmonic mean between precision and recall and is particularly

suitable for unbalanced datasets [8]. In addition, classification accuracy was calculated as the percentage match between the predictions and ground truth. For a detailed error analysis, confusion matrices were generated to provide insights into the systematic misclassifications between humanitarian categories.

The evaluation protocol differs fundamentally from traditional machine-learning approaches in that no train-test split is required. All 18,082 samples from the CrisisMMD dataset were used exclusively for evaluation, with the model receiving no prior training on these or similar data. This true zero-shot evaluation demonstrates the generalization ability of the approach without data-set-specific optimization. The evaluation was performed separately for both classification tasks (informativeness and humanitarian categories) and for both modalities (text and image), enabling differentiated performance analysis.

The experiments were performed on an NVIDIA A100 GPU with 40 GB VRAM in the Google Cloud Platform. During operation, the system requires between 25 and 27 GB of GPU memory, which includes the quantized model version (Q6_K_L), multimodal projector, and processing buffers. The llama.cpp server was configured with a context window of 24,576 tokens, which is significantly higher than the 8192 tokens required for the task at hand, and provides additional flexibility. Parallel processing was set to 16 simultaneous requests, allowing the optimal utilization of GPU resources. The average processing speed is 0.5–1 s per sample with full parallelization.

The implementation was performed in Python 3.10, using established open-source libraries. For model execution, llama-cpp-Python was utilized for efficient quantized inference via a local server with an OpenAI-compatible API endpoint. Image processing was based on Pillow (v10.2) for scaling and format conversion. Pandas (v2.0) and NumPy (v1.24) were used for data processing and metric calculation, whereas Scikit-learn (v1.3) was used to generate confusion matrices and binary F1 score calculations. Parallel processing was implemented with concurrent futures from the Python standard library, with the ThreadPoolExecutor configured with 16 worker threads.

To ensure reproducibility, all relevant implementation details were documented, and the source code was made available via GitHub (<https://github.com/nored/multimodal-disaster-classification>, accessed on 1 February 2026). The model configuration used deterministic settings (temperature = 0) to ensure consistent results across multiple runs.

To assess reproducibility, 500 randomly sampled items (seed = 42) were re-classified in a separate session on identical hardware (NVIDIA A100, 40 GB) using the same model weights, prompt template, JSON schema, and temperature = 0 setting. The retest yielded an 89.8% exact match rate across all five prediction fields (449/500 samples). Mismatch frequency correlated with task granularity: binary informativeness classifications showed 2.2–2.8% instability, four-class damage severity 2.4%, and eight-class humanitarian categorisation 3.4–3.6%. All discrepancies occurred at classification boundaries (e.g., informative/not_informative, adjacent humanitarian categories) with no gross misclassifications observed. These variations are consistent with non-deterministic floating-point reduction order in parallel GPU execution across sessions, a known property of CUDA-based inference even at temperature = 0. The retest notebook and full mismatch report are available in the companion repository.

The llama.cpp server was configured with the chat template “mistral-v7-tekken”, which ensures correct tokenization for the Mistral Small model. All hyperparameters, including the batch size (16), maximum tokens per batch (1024), and image resolution (512 × 512), were hard-coded. The model version used (mistralai_Mistral-Small-3.1-24B-Instruct-2503-Q6_K_L.gguf) and the associated multimodal projector were available via

the Hugging-Face Model Hub. In addition, all classification decisions were logged with timestamps and model outputs to enable full traceability of the results.

4. Results and Discussion

This section presents the results of the zero-shot multimodal classification on the CrisisMMD dataset and discusses its implications for disaster management. Without any task-specific training, the framework achieved binary F1 scores of 0.8415 for text informativeness, 0.8417 for image informativeness, 0.6106 for humanitarian text categorization, and 0.7213 for humanitarian image categorization across 18,082 multimodal samples. The analysis first covers the performance of binary informativeness detection (Section 4.1), followed by a detailed investigation of humanitarian categorization across eight classes (Section 4.2). Next, the generalization ability was evaluated across seven different disaster types (Section 4.3), and the results were classified in the context of traditional approaches (Section 4.4). Finally, practical implications for use in crisis situations (Section 4.5), limitations, and future research directions are discussed (Section 5).

4.1. Informativeness Detection Performance

Informativeness detection is a central component of the framework, and was evaluated on all 18,082 samples in the CrisisMMD dataset. Table 1 lists the overall performance of the zero-shot approach.

Table 1. Overall Classification Performance Summary.

Task	Modality	Binary F1	Weighted F1
Informativeness	Text	0.842	0.789
	Image	0.842	0.831
Humanitarian Category	Text	—	0.611
	Image	—	0.721

The remarkably balanced performance between text (binary F1: 0.842) and image (binary F1: 0.842) demonstrates the robustness of the multimodal approach. Table 2 shows the generalization abilities of the different disaster types.

Table 2. Informativeness Classification Performance by Disaster Type.

Disaster Type	Samples	Text F1	Image F1
California Wildfires	2712	0.8652	0.8690
Hurricane Harvey	5889	0.8400	0.8735
Hurricane Irma	2177	0.8312	0.8178
Hurricane Maria	4630	0.8196	0.8188
Iraq-Iran Earthquake	498	0.8919	0.8630
Mexico Earthquake	1636	0.8616	0.8375
Sri Lanka Floods	540	0.9062	0.8200
Overall	18,082	0.8415	0.8417

The results show consistent performance across all disaster types, with F1 scores between 0.82 and 0.91 for text and between 0.82 and 0.87 for images. Particularly noteworthy is the high performance on smaller datasets such as the Sri Lanka Floods (540 samples, text F1: 0.9062) and the Iraq-Iran Earthquake (498 samples, text F1: 0.8919), which underscores the generalizability of the zero-shot approach.

For the text modality, informativeness classification is the most critical step, as it serves as the primary filter for subsequent sophisticated processing in ODET. A high

precision of 0.8773 ensures that the relevant disaster information is reliably identified. Image classification represents an important extension of this research, with comparably strong results (binary F1: 0.8417), validating the effectiveness of the multimodal approach.

4.2. Humanitarian Category Classification

Humanitarian categorization represents the second level of classification and shows clear differences between modalities. With a weighted F1 score of 0.6106 for text and 0.7213 for images, the framework demonstrated a better performance in the visual analysis of humanitarian content, as summarized in Table 3.

Table 3. Humanitarian Category Classification Performance.

Modality	Weighted F1	Best Category	F1	Worst Category	F1
Text	0.6106	Injured/Dead	0.813	Missing/Found	0.253
Image	0.7213	Not Humanitarian	0.819	Missing/Found	0.105

The superior performance of the imaging modality (F1: 0.7213 vs. 0.6106) can be explained by several factors. Visual content often conveys clearer information about infrastructure damage (image F1: 0.693 vs. text F1: 0.508) and enables more precise identification of non-humanitarian content (image F1: 0.819 vs. text F1: 0.652). This discrepancy suggests that images in crisis situations often contain more direct and less ambiguous information than text descriptions, which are often emotionally or context-dependent.

The analysis of individual categories revealed the systematic strengths and weaknesses of both modalities. Text classification achieved the highest performance for “injured_or_dead_people” (F1:0.813) and “rescue_volunteering_or_donation_effort” (F1: 0.740), demonstrating its effectiveness in recognizing explicit mentions of victims and relief efforts. These categories benefit from clear linguistic markers such as “injured,” “rescue,” or “donation,” which are reliably recognized by the model.

Extreme class imbalances significantly affect performance. The category “missing_or_found_people”, with only 41 text and 14 image samples, shows the weakest performance (text F1: 0.253, image F1: 0.105). By contrast, frequent categories such as “no_humanitarian” (5227 text samples, 8708 image samples) benefit from more robust F1 scores. This imbalance poses a particular challenge for zero-shot approaches because there is no possibility of compensatory weighting during training.

The category “other_relevant_information” shows interesting modality-specific differences: While texts have higher precision (0.76) with low recall (0.41), images show the opposite pattern with higher recall (0.74) and moderate precision (0.51). This suggests that textual descriptions are more precise but more conservative in identifying general disaster information, whereas visual content captures more relevant information but with higher uncertainty.

Despite the class imbalance and challenges of zero-shot classification, the framework achieves practical performance for the most important humanitarian categories. Strong performance in critical categories such as infrastructure damage and rescue operations enables effective filtering and prioritization of disaster content for emergency responders.

4.3. Cross-Disaster Generalization

A key strength of the zero-shot approach is its ability to generalize across different disaster types, without task-specific training. The evaluation across seven different natural disasters demonstrates the robustness and practical applicability of the framework.

The consistency analysis showed remarkable stability across all disaster types. For informativeness detection, the standard deviations were only 0.031 (text) and 0.024 (image),

indicating high generalizability. The F1 scores vary minimally between 0.820–0.906 for text and 0.818–0.874 for images, regardless of the disaster type or dataset size.

The performance on smaller datasets was particularly evident. The Iraq-Iran Earthquake, with only 498 samples, achieved one of the highest text F1 scores (0.892), while the Sri Lanka Floods (540 samples) showed the best text performance (0.906). These results refute the assumption that larger datasets lead to better zero-shot performance, and underscore the independence of this approach from the sample size. Table 4 presents the complete cross-disaster performance analysis.

Table 4. Cross-Disaster Performance Analysis.

Disaster Type	Samples	Informativeness F1		Humanitarian F1 *	
		Text	Image	Text	Image
California Wildfires	2712	0.865	0.869	0.593	0.703
Hurricane Harvey	5889	0.840	0.874	0.608	0.729
Hurricane Irma	2177	0.831	0.818	0.621	0.716
Hurricane Maria	4630	0.820	0.819	0.587	0.714
Iraq-Iran Earthquake	498	0.892	0.863	0.612	0.695
Mexico Earthquake	1636	0.862	0.838	0.601	0.731
Sri Lanka Floods	540	0.906	0.820	0.648	0.762
Mean	-	0.859	0.843	0.610	0.721
Std. Dev.	-	0.031	0.024	0.022	0.022

* Weighted F1 scores for humanitarian categorization.

Humanitarian categorization showed similar consistency, with a standard deviation of 0.022 for both modalities. A consistently higher image performance (mean: 0.721 vs. 0.610) was confirmed across all disaster types, with the Sri Lanka floods again achieving the best results (text: 0.648, image: 0.762).

The analysis of disaster characteristics revealed interesting patterns. Earthquakes (Iraq, Iran, Mexico) consistently showed high informativeness F1 scores, possibly due to clearer visual damage patterns and more accurate reporting. Hurricanes show slightly more variable results, which may reflect the different phases and impacts of long-lasting events.

This cross-disaster generalization validates the practical applicability of the framework. A consistent performance without disaster-specific training means that the system can be deployed immediately for new, unforeseen disasters—a key advantage over traditional, trained approaches that would require annotated data and retraining for each new disaster type.

4.4. Contextualization of Results

The results must be viewed in the context of traditional approaches to disaster classification to fully understand their significance. Conventional machine-learning methods in this field are typically based on supervised learning, which requires significant upfront investment in data collection, annotation, and model training [8].

To contextualize these results, Table 5 provides a direct comparison with both supervised and zero-shot approaches evaluated on the CrisisMMD dataset.

Table 5. Comparison with published approaches on CrisisMMD.

Approach	Type	Info. wF1	Hum. wF1 ^a	Training
<i>Supervised approaches (require annotated training data)</i>				
Ofli et al. [8]	CNN fusion	0.842	0.783	Yes
Abavisani et al. [43]	Cross-attention	—	0.918 ^b	Yes
CrisisBench [44]	Fine-tuned RoBERTa	0.866	0.829	Yes

Table 5. *Cont.*

Approach	Type	Info. wF1	Hum. wF1 ^a	Training
<i>Zero-shot approaches (no task-specific training) ^c</i>				
[45] GPT-4o	Text + image	0.805	0.586 ^d	No
[45] Gemini 1.5 Flash	Text + image	0.796	0.565 ^d	No
[45] Claude 3.5 Sonnet	Text + image	0.795	0.638 ^d	No
Ours (Mistral-Small-3.1)	Text + image	0.789/0.831 ^e	0.611/0.721 ^e	No

^a Weighted F1 for humanitarian categorization. ^b Merged 5-class scheme; not directly comparable to our 8-class setup. ^c Scores from no-event-awareness, with-images configuration on CrisisMMD (Tables III–IV in [45]), matching our experimental setup. ^d 16-class humanitarian scheme; not directly comparable to our 8-class setup. ^e Text/Image weighted F1 scores.

The comparison reveals several noteworthy findings. For informativeness detection, our zero-shot approach achieves weighted F1 scores of 0.789 (text) and 0.831 (image). The image-based score exceeds all three commercial models evaluated by McDaniel et al. [45]—GPT-4o (0.805), Gemini 1.5 Flash (0.796), and Claude 3.5 Sonnet (0.795)—by 2.5–3.5 percentage points. The text-based score falls slightly below these baselines. This modality asymmetry reflects a key strength of our approach: by producing independent text and image classifications, the framework enables responders to leverage whichever modality provides stronger signal for a given disaster type. On binary F1 (informative class only), the commercial models achieve higher scores (0.868–0.871 vs. our 0.842), indicating stronger recall for informative content specifically. However, our model achieves substantially higher macro F1 (0.830 vs. 0.743–0.761 for images), demonstrating more balanced classification across both informative and non-informative content—a property that reduces false alarms in operational settings where precision matters as much as recall. Both zero-shot approaches remain below the supervised state of the art (CrisisBench wF1: 0.866), which is expected given that supervised methods benefit from in-distribution training data.

For humanitarian categorization, our image-based weighted F1 of 0.721 surpasses the best commercial zero-shot result (Claude 3.5 Sonnet: 0.638) by over 8 percentage points. Our text-based score of 0.611 is comparable to these results. This comparison notably favors the commercial models. The evaluation methodology in McDaniel et al. restricts the candidate labels to the subset present in each constituent dataset, effectively simplifying the classification task. In contrast, our framework classifies every sample against the full set of eight humanitarian categories, making our weighted F1 scores a more conservative measure of performance.

A critical contextual factor is the asymmetry in computational resources. McDaniel et al. evaluated commercial models—GPT-4o, Gemini 1.5 Flash, and Claude 3.5 Sonnet—accessed through paid cloud APIs that require transmitting potentially sensitive disaster imagery to external servers. Our framework uses a quantized 24B-parameter open-source model (Mistral-Small-3.1, Q6_K_L) that operates entirely on local infrastructure, requiring only 25–27 GB of GPU memory and no internet connectivity. Despite this resource asymmetry, our image-based informativeness weighted F1 of 0.831 exceeds all three commercial models (0.795–0.805), and our image humanitarian F1 of 0.721 surpasses the best commercial result by over 8 percentage points. On binary F1, the commercial models achieve modestly higher scores for the informative class (0.868–0.871 vs. 0.842), but our model demonstrates substantially more balanced classification (macro F1: 0.830 vs. 0.743–0.761), reducing false alarm rates that are operationally critical in disaster response. That competitive or superior performance is achieved by a model deployable on consumer hardware without data leaving the operator’s infrastructure represents a significant practical advantage for disaster management agencies operating under data sovereignty constraints or

in connectivity-limited environments. The modular architecture of the framework allows straightforward substitution of the underlying model, and future work should benchmark additional vision-language models, including GPT-4V, Gemini, and LLaVA, to establish comprehensive performance baselines across model families.

These comparisons reveal a fundamental trade-off in crisis classification systems. Supervised approaches achieve higher absolute performance when in-distribution training data is available, as demonstrated by CrisisBench (wF1: 0.866) compared to our best zero-shot result (wF1: 0.831). However, this accuracy advantage must be weighed against substantial operational costs: supervised systems require days to weeks for data annotation and training, are constrained to disaster types present in their training data, and demand continuous retraining to counter concept drift. Zero-shot approaches accept a modest accuracy margin in exchange for immediate deployability, inherent generalizability to novel disaster types, and elimination of ongoing maintenance.

Our zero-shot results of binary F1 scores above 0.84 for informativeness detection without any task-specific training represent a paradigm shift. This performance, which is comparable to that of trained baseline models, was achieved solely through careful prompt engineering and leveraging the inherent capabilities of modern vision-language models. The fact that these results were achieved without a single annotated training data point from the CrisisMMD dataset demonstrates the evolution of AI technology since its introduction.

The real-world relevance of this approach manifests in several critical advantages. Immediate deployability in new disasters eliminates the typical latency between the event onset and a functional classification system. While traditional approaches require days or weeks for data collection and training, the framework can be operational within the minutes of a new disaster. This speed can be lifesaving during the critical first hours of a disaster.

The sustainability of the zero-shot approach is evident in the elimination of the ongoing maintenance requirements. Traditional models suffer from concept drift and must be retrained regularly using new data to maintain their performance. On the other hand, the framework requires no retraining and remains stable over time because it is based on fundamental linguistic and visual concepts encoded in the base models.

The demonstrated generalizability across the seven different disaster types with consistent F1 scores (standard deviation < 0.031) validates the robustness of the approach. This consistency without disaster-specific adjustments means that emergency managers can rely on predictable system performance, regardless of the type of crisis. The ability to handle both known and completely new disaster types makes the system future-proof of changing climate patterns and emerging threats.

This contextualization shows that zero-shot multimodal classification is not just a technical curiosity but offers a viable and, in many respects, superior alternative to traditional approaches to disaster management. The combination of immediate operational readiness, consistent performance, and minimal maintenance requirements makes this methodology an important building block for modern crisis response systems.

4.5. Practical Implications

This study answers a fundamental research question: Is zero-shot classification practical for multimodal disaster analysis, and what implementation effort is required? The results clearly demonstrate that modern vision-language models can achieve practical binary F1 scores above 0.84 without task-specific training.

The implementation effort for practical deployment was moderate. The core components—prompt engineering, JSON schema definition, and API integration—required approximately two weeks of development time. The critical steps include iterative op-

timization of the prompts for consistent classification results, definition of structured output formats for both modalities, and implementation of robust error handling for batch processing. None of these tasks required specialized machine learning or complex training procedures.

Practical applicability is evident in several respects. First, the zero-shot approach eliminates the typical latency between problem identification and ready-to-use systems. Although traditional approaches require weeks for data collection, annotation, and training, the system was ready to be used immediately after the initial prompt development. Second, maintenance efforts are drastically reduced: no retraining cycles, no annotation of new data, and no hyperparameter optimization.

The extension of the ODET to include multimodal capabilities validates the viability of the zero-shot paradigm. With minimal adjustments, mainly the addition of image processing logic and modality-specific prompts, the existing text-based pipeline can be extended to include visual analysis. This modularity confirms that zero-shot approaches are suitable not only for isolated proof-of-concepts but also for complex, production-ready systems.

Particularly insightful is the realization that the main work lies not in model adaptation but in system integration. The challenges primarily concerned practical aspects, such as batch processing, memory management, and API stability—all solvable engineering problems without AI-specific complexity. This shift from ML expertise to software engineering makes technology accessible to a broader range of organizations.

The generalizability across the seven disaster types without performance degradation confirms the robustness of the approach. Organizations can be confident that, once developed, zero-shot systems will continue to work for future, unforeseen events. This future-proofing justifies the initial development investment and makes the zero-shot classification a sustainable technology choice for disaster management.

Our experience shows that the barrier to the practical use of zero-shot multimodal classification is primarily conceptual and not technical. Once organizations accept the paradigm of the training-free approach, its implementation is straightforward, and the results are immediately usable. This finding should encourage more research groups and practitioners to explore zero-shot methods for specific cases.

5. Limitations and Future Work

Despite these promising results, the zero-shot approach has several limitations that should be addressed in future research. The pronounced class imbalance in the CrisisMMD dataset significantly affects the performance, especially for rare categories. The category “missing_or_found_people,” with only 41 text and 14 image samples, achieves F1 scores of only 0.253 (text) and 0.105 (image), while frequent categories such as “not_humanitarian,” with over 5000 samples, achieved robust scores above 0.65. This discrepancy is particularly problematic because rare categories such as missing persons can be critical in crisis situations.

The zero-shot nature of this approach exacerbates the imbalance problem because there is no possibility of adjusting the weighting during training. However, these results must be interpreted with caution given the extremely limited sample sizes for this category. With only 14 image annotations, individual misclassifications have a disproportionate impact on evaluation metrics, making reliable performance assessment statistically challenging. To our knowledge, no existing multimodal crisis benchmark provides sufficient annotations for missing persons detection, highlighting a significant gap in available evaluation resources. Future research should evaluate this and other rare but critical categories on larger, dedicated datasets with more balanced class distributions to determine whether the

low performance reflects a fundamental classification challenge or primarily a measurement artifact of insufficient evaluation data.

More broadly, the CrisisMMD dataset is limited to four types of natural disasters: hurricanes, earthquakes, wildfires, and floods. Other disaster categories such as extreme heat events, severe storms, cold waves, and industrial accidents including chemical or nuclear incidents are not represented. This observation extends beyond our work: the field of multimodal crisis informatics currently lacks benchmark datasets for these disaster types entirely [18]. The creation of multimodal crisis datasets covering a broader spectrum of both natural and human-made disasters, with balanced annotations across all humanitarian categories, remains an urgent need for the research community and a prerequisite for comprehensive evaluation of any classification framework. While the consistent cross-disaster generalization observed across the seven evaluated events (standard deviation below 0.031) suggests that the zero-shot approach is well positioned to handle novel disaster types, this assumption remains empirically unvalidated for disaster categories with fundamentally different visual and textual characteristics.

Additionally, the framework inherits potential biases from the pre-trained vision-language model. These models are predominantly trained on English-language, Western-centric web data, which may result in reduced classification accuracy for disaster content from underrepresented geographic regions, non-Western building styles, or non-English social media posts. In operational deployment within the ODET pipeline [31], these risks are partially mitigated at multiple stages: topical clustering of large post volumes provides statistical robustness against individual misclassifications, while retrieval-augmented generation and knowledge graph integration reduce hallucination risks during report synthesis. Nevertheless, hallucination remains a fundamental and unsolved challenge inherent to current large language model architectures, and its complete elimination cannot be guaranteed at any pipeline stage. The extent to which pre-trained model biases and hallucination artifacts affect classification and reporting accuracy in diverse global disaster contexts remains an important direction for future research.

The CrisisMMD dataset includes damage severity annotations for a subset of 3526 images. However, these annotations were excluded from the present evaluation owing to significant inconsistencies in the labeling scheme. The three-level severity scale (“little or no damage”, “mild damage”, “severe damage”) exhibited substantial inter-annotator disagreement in the original dataset, and the boundaries between severity levels are inherently subjective. Initial exploratory analysis confirmed that zero-shot classification of gradual severity levels poses challenges distinct from binary or categorical classification. Future research should develop specialized prompting strategies for ordinal classification tasks, possibly incorporating reference images or detailed visual damage criteria to anchor severity judgments.

The current implementation produces text and image classifications in separate output fields, without explicitly exploiting cross-modal consistency checks. Future versions should integrate cross-modal attention mechanisms to identify conflicting or complementary information between the text and images. A tweet claiming “everything is fine” while the associated image shows that severe damage should be treated differently than consistent text-image pairs.

Integration into a complete ODET pipeline remains a challenge. Although this study demonstrates the feasibility of multimodal classification, seamless integration into the existing components—semantic clustering, knowledge graph generation, and report synthesis—requires further development. In particular, the fusion of multimodal information with coherent narratives for situational awareness reports requires sophisticated aggregation strategies.

The scaling aspects were theoretically addressed. While the system works with 16 batches, the modern context windows of 24,576 tokens could theoretically process hundreds of samples simultaneously. However, the practical implementation of such mega-batches requires careful memory management and new prompting paradigms that can efficiently handle extreme context lengths.

Language dependency is another limitation of this study. Although image analysis is language-independent, the prompts were based on English instruction. Multilingual disaster response requires either translated prompts or multilingual models. Evaluating the cross-lingual transfer capabilities of zero-shot approaches in disaster contexts remains an important research area.

Finally, this study did not address the ethical implications of automated disaster classifications. False negatives in critical categories such as “injured_or_dead_people” could have real-world consequences. Future studies should integrate uncertainty quantification and develop clear protocols for human oversight in critical decisions.

6. Conclusions

This study successfully demonstrates the practicality of zero-shot multimodal classification for disaster management by extending the ODET framework with image analysis capabilities. With binary F1 scores of 0.8415 (text) and 0.8417 (image) for informativeness detection and 0.6106 (text) and 0.7213 (image) for humanitarian categorization, the training-free approach achieves performance levels that are relevant for practical applications.

The core contributions of this study comprise three key aspects. First, it was proven that modern vision-language models enable effective disaster classification without task-specific training. Second, consistent generalization was demonstrated across seven different disaster types with minimal performance variance (standard deviation < 0.031). Third, it is shown that the implementation effort for zero-shot systems lies primarily in prompt engineering and system integration but not in complex ML processes.

The practical significance of these results cannot be overstated. In a world of increasing natural disasters and limited resources for data annotation, zero-shot classification offers a viable path for immediate crisis response. The elimination of training time means that responders can have functional analysis systems within minutes of a new disaster. This speed can be lifesaving during the critical first hour.

The successful integration of multimodal capabilities into ODET is an important evolutionary step. Although text-based analysis already provides valuable insights, image processing adds an essential dimension. In an increasingly visual social media landscape, the ability to understand and integrate both modalities is not optional but necessary for comprehensive situational awareness.

Our experience validates zero-shot approaches as mature technologies for productive systems. The key finding that implementation hurdles are primarily engineering rather than algorithmic in nature should encourage more organizations to adopt this technology. With a development time of about two weeks for a working system, the investment is manageable, while the benefits—immediate, training-free, and generalizable classification—are significant.

Future research should address the identified limitations, particularly the class imbalance problem and the integration of advanced multimodal fusion mechanisms. Further development of a fully integrated multimodal ODET system that generates coherent reports from text and image information remains an exciting challenge.

This study contributes to the growing evidence that the machine-learning paradigm is undergoing a fundamental shift. The shift is away from data-intensive, task-specific models toward flexible, generalist systems that demonstrate human-like adaptability. For time-

critical applications, such as disaster management, this shift is not only technologically interesting but also operationally transformative. Zero-shot multimodal classification is no longer a vision of the future; it is a practical reality, ready for deployment in the next crisis.

Author Contributions: Conceptualization, F.S.; methodology, F.S. and K.D.S.; validation, F.S., K.D.S., D.A.A., K.B., N.S. and R.C.; writing—original draft, F.S. and K.D.S.; writing—review and editing, D.A.A., K.B., N.S., R.C. and V.D.; supervision, D.A.A., R.C. and V.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: <https://crisisnlp.qcri.org/crisismmd> and here: <https://github.com/nored/multimodal-disaster-classification> (accessed on 1 February 2026).

Acknowledgments: This work was supported partially by the European Union in the framework of ERASMUS MUNDUS, Project CyberMACS (Project #101082683).

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
BLIP	Bootstrapping Language-Image Pre-training
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CRED	Centre for Research on the Epidemiology of Disasters
CSV	Comma-Separated Values
DAVI	Disaster Assessment with Vision Foundation Model
ECG	Electrocardiogram
F1	F1 Score (harmonic mean of precision and recall)
GB	Gigabyte
GGUF	GPT-Generated Unified Format
GPU	Graphics Processing Unit
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
LLM	Large Language Model
LVL	Large Vision-Language Model
MERL	Multimodal ECG Representation Learning
ML	Machine Learning
MLLM	Multimodal Large Language Model
NLP	Natural Language Processing
ODET	Open-Source Intelligence Disaster Event Tracker
QCRI	Qatar Computing Research Institute
RGB	Red Green Blue
RGBA	Red Green Blue Alpha
UNDRR	United Nations Office for Disaster Risk Reduction
URL	Uniform Resource Locator
VERO	Verification and Zero-shot Feedback Acquisition
VRAM	Video Random Access Memory
WMO	World Meteorological Organization

References

1. Karimiziarani, M. Social Media Analytics in Disaster Response: A Comprehensive Review. *arXiv* **2023**, arXiv:2307.04046. [\[CrossRef\]](#)
2. Alharbi, K.; Haq, M.A. Enhancing Disaster Response and Public Safety with Advanced Social Media Analytics and Natural Language Processing. *Eng. Technol. Appl. Sci. Res.* **2024**, *14*, 14212–14218. [\[CrossRef\]](#)
3. He, C.; Hu, D. Social Media Analytics for Disaster Response: Classification and Geospatial Visualization Framework. *Appl. Sci.* **2025**, *15*, 4330. [\[CrossRef\]](#)
4. Gupta, T.; Roy, S. Handcrafted Features Based Analysis of Social Media Images for Disaster Response. In *Proceedings of the 2023 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, Cosenza, Italy, 13–15 September 2023; IEEE: New York, NY, USA, 2023; pp. 1–6. [\[CrossRef\]](#)
5. Ponce-López, V.; Spataru, C. Social Media Data Analysis Framework for Disaster Response. *Discov. Artif. Intell.* **2022**, *2*, 10. [\[CrossRef\]](#)
6. Basit, M.; Alam, B.; Fatima, Z.; Shaikh, S. Natural Disaster Tweets Classification Using Multimodal Data. In *Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 6–10 December 2023; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 7584–7594. [\[CrossRef\]](#)
7. Firmansyah, H.B.; Fernandez-Marquez, J.L.; Cerquides, J.; Lorini, V.; Alberto Bono, C.; Pernici, B. Enhancing Disaster Response with Automated Text Information Extraction from Social Media Images. In *Proceedings of the 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)*, Athens, Greece, 17–20 July 2023; IEEE: New York, NY, USA, 2023; pp. 71–78. [\[CrossRef\]](#)
8. Ofli, F.; Alam, F.; Imran, M. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. In *Proceedings of the 17th International Conference on Information Systems for Crisis Response and Management ISCRAM*, Blacksburg, VA, USA, 24–27 May 2020; ISCRAM: Brussels, Belgium, 2020.
9. Gao, S.; Yang, T.; Xu, Y.; Mou, N.; Wang, X.; Huang, H. Enhancing Disaster Situation Awareness Through Multimodal Social Media Data: Evidence from Typhoon Haikui. *Appl. Sci.* **2025**, *15*, 465. [\[CrossRef\]](#)
10. Temkar, R. AI-Powered Multimodal Disaster Response Enhancement using Social Media Streams. *Commun. Appl. Nonlinear Anal.* **2024**, *32*, 198–211. [\[CrossRef\]](#)
11. Hanny, D.; Resch, B. Multimodal Geo-Information Extraction from Social Media for Supporting Decision-Making in Disaster Management. *Agil. Gisci. Ser.* **2024**, *5*, 28. [\[CrossRef\]](#)
12. Firmansyah, H.B.; Lorini, V.; Mulayim, M.O.; Gomes, J.; Fernandez-Marquez, J.L. Improving Social Media Geolocation for Disaster Response by Using Text From Images and ChatGPT. In *Proceedings of the 2024 11th Multidisciplinary International Social Networks Conference*, Bali, Indonesia, 21–23 August 2024; MISNC: New York, NY, USA, 2024; pp. 67–72. [\[CrossRef\]](#)
13. Chen, Z.; Shamsabadi, E.A.; Jiang, S.; Shen, L.; Dias-da Costa, D. Integration of Large Vision Language Models for Efficient Post-disaster Damage Assessment and Reporting. *arXiv* **2024**, arXiv:2411.01511. [\[CrossRef\]](#)
14. Lei, Z.; Dong, Y.; Li, W.; Ding, R.; Wang, Q.; Li, J. Harnessing Large Language Models for Disaster Management: A Survey. *arXiv* **2025**, arXiv:2501.06932. [\[CrossRef\]](#)
15. Kadiyala, L.A.; Mermer, O.; Samuel, D.J.; Sermet, Y.; Demir, I. The Implementation of Multimodal Large Language Models for Hydrological Applications: A Comparative Study of GPT-4 Vision, Gemini, LLaVa, and Multimodal-GPT. *Hydrology* **2024**, *11*, 148. [\[CrossRef\]](#)
16. Imran, M.; Ofli, F.; Caragea, D.; Torralba, A. Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. *Inf. Process. Manag.* **2020**, *57*, 102261. [\[CrossRef\]](#)
17. Ahn, K.; Han, S.; Park, S.; Kim, J.; Park, S.; Cha, M. Generalizable Disaster Damage Assessment via Change Detection with Vision Foundation Model. *arXiv* **2024**, arXiv:2406.08020. [\[CrossRef\]](#)
18. Algiriyage, N.; Prasanna, R.; Stock, K.; Doyle, E.E.H.; Johnston, D. Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review. *SN Comput. Sci.* **2021**, *3*, 92. [\[CrossRef\]](#)
19. Zou, Z.; Gan, H.; Huang, Q.; Cai, T.; Cao, K. Disaster Image Classification by Fusing Multimodal Social Media Data. *Isprs Int. J. -Geo-Inf.* **2021**, *10*, 636. [\[CrossRef\]](#)
20. Liu, C.; Wan, Z.; Ouyang, C.; Shah, A.; Bai, W.; Arcucci, R. Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement. *arXiv* **2024**, arXiv:2403.06659. [\[CrossRef\]](#)
21. Abdelhamed, A.; Afifi, M.; Go, A. What Do You See? Enhancing Zero-Shot Image Classification with Multimodal Large Language Models. *arXiv* **2024**, arXiv:2405.15668. [\[CrossRef\]](#)

22. Wang, Q.; Liu, G.; Wang, B. CapS-Adapter: Caption-based MultiModal Adapter in Zero-Shot Classification. In Proceedings of the 32nd ACM International Conference on Multimedia, ACM, MM '24, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 5442–5450. [\[CrossRef\]](#)
23. Udandara, V.; Prabhu, A.; Ghosh, A.; Sharma, Y.; Torr, P.H.S.; Bibi, A.; Albanie, S.; Bethge, M. No “Zero-Shot” Without Exponential Data: Pretraining Concept Frequency Determines Multimodal Model Performance. *arXiv* **2024**, arXiv:2404.04125. [\[CrossRef\]](#)
24. Chen, X.; Zheng, M.; Liu, B.; Zhang, Y.; Chen, C.; Zhang, P. Large multimodal model for zero-shot scene classification of high spatial resolution remote sensing images. *Remote Sens. Lett.* **2025**, *16*, 449–459. [\[CrossRef\]](#)
25. Islam, A.; Biswas, M.R.; Zaghoulani, W.; Belhaouari, S.B.; Shah, Z. Pushing Boundaries: Exploring Zero Shot Object Classification with Large Multimodal Models. In *Proceedings of the 2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS), Abu Dhabi, United Arab Emirates, 21–24 November 2023*; IEEE: New York, NY, USA, 2023; pp. 1–5. [\[CrossRef\]](#)
26. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597. [\[CrossRef\]](#)
27. Hughes, A.L.; Clark, H. Seeing the Storm: Leveraging Multimodal LLMs for Disaster Social Media Video Filtering. In Proceedings of the International ISCRAM Conference, Halifax, NS, Canada, 18–21 May 2025. [\[CrossRef\]](#)
28. Hickmon, J. Multimodal Ensembling for Zero-Shot Image Classification. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 23747–23749. [\[CrossRef\]](#)
29. Ravi, R.; Kannappan, J. ZeroEVNet: A multimodal zero-shot learning framework for scalable emergency vehicle detection. *Expert Syst. Appl.* **2025**, *275*, 126934. [\[CrossRef\]](#)
30. Risso, M.; Goffi, A.; Motetti, B.A.; Burrello, A.; Bove, J.B.; Macii, E.; Poncino, M.; Pagliari, D.J.; Maffei, G. Building Damage Assessment in Conflict Zones: A Deep Learning Approach Using Geospatial Sub-Meter Resolution Data. In *Proceedings of the 2025 IEEE 6th International Conference on Image Processing, Applications and Systems (IPAS), Lyon, France, 9–11 January 2025*; IEEE: New York, NY, USA, 2025; pp. 1–6. [\[CrossRef\]](#)
31. Schwarz, K.; Bollens, K.; Arias Aranda, D.; Hartmann, M. AI-Enhanced Disaster Management: A Modular OSINT System for Rapid Automated Reporting. *Appl. Sci.* **2024**, *14*, 11165. [\[CrossRef\]](#)
32. Ponce-López, V.; Spataru, C. Behaviour in social media for floods and heat waves in disaster response via Artificial Intelligence. *arXiv* **2022**, arXiv:2203.08753. [\[CrossRef\]](#)
33. Aljurbua, R.; Alshehri, J.; Alharbi, A.; Power, W.; Obradovic, Z. Social Media Sensors for Weather-Caused Outage Prediction Based on Spatio-Temporal Multiplex Network Representation. *IEEE Access* **2023**, *11*, 125883–125896. [\[CrossRef\]](#)
34. Otudi, H.; Gupta, S.; Albarakati, N.; Obradovic, Z. Classifying Severe Weather Events by Utilizing Social Sensor Data and Social Network Analysis. In *Proceedings of the 2023 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM '23, Kusadasi, Türkiye, 6–9 November 2023*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 64–71. [\[CrossRef\]](#)
35. Sun, K.; Wu, H.; Shi, B.; Mensah, S.; Liu, P.; Dong, B. VERO: Verification and Zero-Shot Feedback Acquisition for Few-Shot Multimodal Aspect-Level Sentiment Classification. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 25210–25218. [\[CrossRef\]](#)
36. Abu Tami, M.; Ashqar, H.I.; Elhenawy, M.; Glaser, S.; Rakotonirainy, A. Using Multimodal Large Language Models (MLLMs) for Automated Detection of Traffic Safety-Critical Events. *Vehicles* **2024**, *6*, 1571–1590. [\[CrossRef\]](#)
37. Goecks, V.G.; Waytowich, N.R. DisasterResponseGPT: Large Language Models for Accelerated Plan of Action Development in Disaster Response Scenarios. *arXiv* **2023**, arXiv:2306.17271. [\[CrossRef\]](#)
38. Kumbam, P.R.; Vejre, K.M. FloodLense: A Framework for ChatGPT-based Real-time Flood Detection. *arXiv* **2024**, arXiv:2401.15501. [\[CrossRef\]](#)
39. Kalluri, T.; Lee, J.; Sohn, K.; Singla, S.; Chandraker, M.; Xu, J.; Liu, J. Robust Disaster Assessment from Aerial Imagery Using Text-to-Image Synthetic Data. In *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 17–18 June 2024*; IEEE: New York, NY, USA, 2024; pp. 7449–7459. [\[CrossRef\]](#)
40. Rawat, R. DisasterQA: A Benchmark for Assessing the performance of LLMs in Disaster Response. *arXiv* **2024**, arXiv:2410.20707. [\[CrossRef\]](#)
41. Hackathon, M.S. Leveraging AI for Natural Disaster Management: Takeaways From The Moroccan Earthquake. *arXiv* **2023**, arXiv:2311.08999. [\[CrossRef\]](#)
42. Alam, F.; Ofli, F.; Imran, M. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM), Palo Alto, CA, USA, 25–28 June 2018.
43. Abavisani, M.; Wu, L.; Hu, S.; Tetreault, J.; Jaimes, A. Multimodal Categorization of Crisis Events in Social Media. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020*; IEEE: New York, NY, USA, 2020; pp. 14667–14677. [\[CrossRef\]](#)

44. Alam, F.; Sajjad, H.; Imran, M.; Ofli, F. CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. *Proc. Int. AAAI Conf. Web Soc. Media* **2021**, *15*, 923–932. [[CrossRef](#)]
45. McDaniel, E.L.; Scheele, S.; Liu, J. Zero-Shot Classification of Crisis Tweets Using Instruction-Finetuned Large Language Models. In *Proceedings of the 2024 IEEE International Humanitarian Technologies Conference (IHTC), Bari, Italy, 27–30 November 2024*; IEEE: New York, NY, USA, 2024; pp. 1–7. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.