International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 5, Number: 1-1, Pages: 60- 71, Year: 2026

Multimodal Transfer Learning for Crisis Response: A Unified DeBERTa–ResNet Framework on CrisisMMD

Haffiz-Ud-Din*, Musharraf Ali, Anzal Hussain, Akmal Hussain, Syed Qamar Abbas Kazmi, Hannan Hussain, Syed Hassan Abbas and Sohail Ahmed

Department of Research and Development, Gilgit-Baltistan First Research Center, Binary HUB, Gilgit-Baltistan, Pakistan
* Corresponding author: Haffiz-Ud-Din (Email: haffizuddin7899@gmail.com)

Received: 07/08/2026, Revised: 17/10/2026, Accepted: 03/11/2026

Abstract*— Social media platforms generate large volumes of multimodal content during natural disasters that can support situational awareness and humanitarian decision-making. However, effectively analyzing both textual and visual information remains challenging due to weak cross-modal alignment, severe class imbalance, noisy user-generated data, and limited labeled samples. This paper presents a unified multimodal transfer-learning framework that combines a frozen DeBERTa-v3-base text encoder with a frozen ResNet-50 image encoder through a simple late-fusion strategy for crisis-related classification. The architecture projects the unimodal features into a common 256-dimensional space, concatenates them, and feeds the resulting representation into a lightweight multi-layer perceptron classifier. The framework is evaluated on all three official tasks of the CrisisMMD v2.0 dataset. Experimental results show an accuracy of 0.76 (macro-F1 = 0.74) on the informativeness task, 0.43 (macro-F1 = 0.39) on the humanitarian categories task, and 0.69 (macro-F1 = 0.64) on the damage severity task. The model performs relatively well on classes with distinctive multimodal signatures, while residual confusion remains on semantically overlapping or visually subtle categories. For the damage severity task, which is primarily image-based, the multimodal model achieves essentially the same performance as an image-only configuration, indicating limited contribution from text. The study also acknowledges that formal text-only and image-only ablation experiments under the same protocol were not conducted, so the precise benefit of fusion cannot be rigorously quantified. Overall, the work demonstrates that a computationally efficient frozen late-fusion architecture can serve as a practical and reproducible reference point for multimodal crisis informatics under the resource constraints typical of real-world disaster response.

***Index Terms*— Multimodal deep learning, transfer learning, crisis informatics, DeBERTa, ResNet, social media, disaster response, CrisisMMD**

I. Introduction

Natural disasters and humanitarian crises generate massive volumes of real-time information on social media platforms, particularly Twitter (now X). Citizens, first responders, and affected populations post textual messages accompanied by images that describe injured or deceased people, infrastructure damage, rescue efforts, and other critical situational details. When processed effectively, this multimodal content can significantly enhance situational awareness and support timely decision-making by humanitarian organizations [1], [2]. Traditional approaches that rely solely on text or solely on images often miss complementary information present in the other modality. Consequently, multimodal deep learning techniques that jointly analyze text and visual content have emerged as a promising direction for crisis informatics research.

This paper presents a unified multimodal transfer-learning framework that combines a powerful pretrained language model (DeBERTa-v3-base [3]) with a strong visual backbone (ResNet-50 [4]) through late fusion. The framework is systematically evaluated on all three major annotation tasks of the CrisisMMD dataset: informativeness classification, humanitarian category classification, and damage severity assessment. By freezing the pretrained backbones and training only lightweight classification heads, the proposed approach demonstrates an efficient and effective way to leverage large-scale pretrained models for crisis-related multimodal classification.

A. Background & Motivation

Social media has become an indispensable source of information during disasters. Studies have repeatedly shown that tweets and associated images can provide early indicators of damage, needs, and emerging risks that are not immediately available through traditional channels [1], [5], [6, 7]. The CrisisMMD dataset, released by Alam et al. [1], was the first large-scale publicly available multimodal Twitter dataset collected from seven major natural disasters in 2017. It provides human annotations for three complementary tasks:

i. Informativeness (whether a post is useful for humanitarian response),
ii. Humanitarian categories (affected individuals, infrastructure damage, rescue efforts, etc.),
iii. Damage severity assessment (severe, mild, or little/no damage).

This work is licensed under a Creative Commons AttributionShareAlike4.0 International License, which permits Unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

Subsequent work by Ofli et al. [2] demonstrated that multimodal models can outperform unimodal baselines on these tasks. However, most existing multimodal approaches still rely on relatively older CNN-based text encoders or train models from scratch / with limited transfer learning. The rapid progress in pretrained transformer language models (such as DeBERTa [3]) and residual visual networks (ResNet [4]) opens new opportunities for stronger and more sample-efficient multimodal architectures.

The motivation of this research is therefore twofold. First, to investigate whether modern pretrained models (DeBERTa for text and ResNet-50 for images) can be effectively combined through a simple yet powerful late-fusion strategy for crisis-related classification. Second, to provide a comprehensive multi-task evaluation on the entire CrisisMMD benchmark under a consistent experimental setup, thereby offering a strong and reproducible baseline for future multimodal crisis informatics research.

B. Challenges of Multimodal Social Media Data in Crises

Despite its potential, multimodal analysis of crisis-related social media data presents several significant challenges:

1. **Weak Alignment Between Modalities** Text and image in the same post frequently convey complementary or even partially conflicting information. A tweet may describe severe damage while the attached image shows only mild destruction, or vice versa [2]. This weak cross-modal alignment makes simple early-fusion strategies less effective.
2. **Severe Class Imbalance** Humanitarian categories such as "missing or found people" or "vehicle damage" contain very few samples compared with dominant classes such as "not humanitarian" or "other relevant information" [1]. Similarly, damage severity labels are highly skewed toward severe damage. Without proper balancing or class-weighted losses, models tend to ignore minority classes that are often the most critical for response.
3. **Noisy and Incomplete Data** User-generated content frequently contains informal language, abbreviations, misspellings, hashtags, and low-quality or irrelevant images. Missing images or corrupted files further complicate training pipelines.
4. **Limited Labeled Data and Domain Shift** Although CrisisMMD is one of the largest multimodal crisis datasets, the absolute number of labeled samples remains modest compared with general-domain vision-language datasets. Models pretrained on general web data may suffer from domain shift when applied to disaster imagery and crisis-specific language.
5. **Computational and Deployment Constraints** Humanitarian organizations often require models that can be trained or adapted quickly on limited hardware and that remain computationally efficient for near real-time inference during unfolding crises.

These challenges motivate the design choices adopted in this work: the use of strong pretrained backbones with frozen weights (to reduce the need for large labeled data and computational resources), late fusion (to handle weak modality alignment), and careful data balancing strategies.

C. Research Objectives

The primary goal of this research is to develop and evaluate a unified multimodal transfer-learning framework that effectively combines textual and visual information from social media posts for crisis response applications. Specifically, the study pursues the following objectives:

i. To design a late-fusion multimodal architecture that integrates a state-of-the-art pretrained language model (DeBERTa-v3-base) with a robust visual backbone (ResNet-50) while keeping the pretrained parameters frozen, thereby enabling efficient transfer learning on limited labeled crisis data.
ii. To systematically evaluate the proposed framework on all three core annotation tasks of the CrisisMMD dataset-informativeness classification, humanitarian category classification, and damage severity assessment-under a consistent experimental protocol.
iii. To investigate the impact of data balancing strategies and the effectiveness of late fusion in handling the inherent challenges of weakly aligned multimodal crisis data.
iv. To provide a strong, reproducible multimodal baseline that can serve as a reference point for future research in multimodal crisis informatics.

D. Contributions

The main contributions of this work are summarized as follows:

i. We propose a simple and computationally efficient late-fusion multimodal architecture that combines frozen DeBERTa-v3-base text representations with frozen ResNet-50 image representations for crisis-related classification. By keeping the pretrained backbones frozen and training only lightweight projection and classification layers, the framework remains practical under the resource constraints typically faced in real-world disaster response settings.
ii. We evaluate the proposed framework consistently across all three official tasks of the CrisisMMD v2.0 benchmark — informativeness classification, humanitarian category classification, and damage severity assessment — under the same experimental protocol. This provides a unified view of how a single frozen late-fusion design performs on tasks of increasing complexity and class imbalance.
iii. We show that combining strong pretrained unimodal features with training-time class balancing yields reasonable performance on the highly imbalanced humanitarian and damage severity tasks, without requiring full fine-tuning of large backbone networks. The model performs particularly well on classes with distinctive multimodal signatures, while residual confusion remains on semantically overlapping or visually subtle categories.
iv. We provide a transparent and reproducible experimental pipeline (including data preprocessing, balancing strategy, model configuration, and

evaluation protocol) that can serve as a practical baseline for future comparative studies in multimodal crisis informatics.

E. Paper Organization

The remainder of this paper is organized as follows. Section II reviews related work on text-based, image-based, and multimodal approaches for crisis informatics, as well as recent advances in transfer learning. Section III describes the CrisisMMD dataset, the three classification tasks, and the preprocessing and balancing procedures employed in this study. Section IV presents the proposed multimodal architecture, including the text encoder, image encoder, fusion strategy, and training details. Section V outlines the experimental setup, evaluation metrics, and implementation specifics. Section VI reports and discusses the results obtained on the three tasks. Finally, Section VII concludes the paper and outlines directions for future research.

II. Related Work

Research on the automatic analysis of social media content for disaster response has evolved from purely text-based methods to image-based approaches and, more recently, to multimodal techniques. This section reviews the most relevant prior work in each of these directions, with particular emphasis on studies that utilize the CrisisMMD dataset or address similar crisis-related classification tasks.

A. Text-only Crisis Classification

Early efforts in crisis informatics primarily focused on textual content. Researchers developed systems to classify tweets according to informativeness, humanitarian categories, urgency, and other actionable classes. Classical machine learning approaches using bag-of-words, TF-IDF, and handcrafted linguistic features were widely employed before the deep learning era. With the advent of neural networks, convolutional and recurrent architectures, followed by transformer-based models, significantly improved performance on crisis tweet classification tasks.

Several benchmark datasets, such as CrisisLex and CrisisNLP collections, facilitated progress in this area. On the CrisisMMD dataset specifically, text-only baselines have been reported in the original dataset paper [1] and in subsequent multimodal studies [2]. These works demonstrated that strong text encoders can achieve reasonable performance on informativeness and humanitarian category classification. However, purely textual models often fail to capture critical visual evidence of damage severity or the physical context of an event, limiting their usefulness for certain operational tasks such as damage assessment.

Recent advances in pretrained language models, particularly DeBERTa [3], have shown superior performance over earlier BERT-style models on a wide range of natural language understanding benchmarks. Their application to the crisis domain remains relatively limited, creating an opportunity to explore stronger text representations within multimodal frameworks.

B. Image-only Damage Assessment

Parallel to text-based research, a growing body of work has focused on analyzing images posted on social media during disasters. The primary objective of these studies is typically damage severity assessment or the detection of infrastructure destruction. Convolutional neural networks, especially residual architectures such as ResNet [4], have become the dominant approach due to their strong feature extraction capabilities.

Nguyen et al. [7] conducted one of the early systematic studies on damage assessment from social media imagery and showed that domain-specific fine-tuning of deep CNNs outperforms traditional bag-of-visual-words methods. Mozannar et al. [6] further explored multimodal combinations but also reported competitive results using image-only models. Alam et al. [8] proposed Image4Act, a system specifically designed for real-time processing of social media images for disaster response.

On the CrisisMMD dataset, the damage severity task is defined exclusively on images [1], [2]. Image-only models using ResNet or similar backbones have been used as strong baselines. While these models perform well when clear visual evidence of damage is present, they struggle with images that lack obvious destruction cues or when the damage is better described in the accompanying text. This limitation further motivates the integration of both modalities [9].

C. Multimodal Approaches for Disaster Response

Recognizing the complementary nature of text and images, researchers have increasingly turned to multimodal deep learning for crisis informatics. Early multimodal systems typically extracted features from each modality independently using CNNs for images and CNNs or RNNs for text, followed by concatenation or simple fusion layers. Mozannar et al. [6] proposed one of the first multimodal deep learning frameworks for damage identification in social media posts, combining pretrained unimodal networks before a final classifier.

Ofli et al. [2] presented a more systematic multimodal study on the CrisisMMD dataset. They designed a modality-agnostic shared representation architecture based on convolutional networks and demonstrated that multimodal models consistently outperform strong text-only and image-only baselines on both informativeness and humanitarian category classification tasks. Their work remains one of the most cited multimodal baselines on CrisisMMD.

More recent studies have explored advanced fusion strategies, including cross-modal attention and graph-based approaches [10], [11]. While these methods often achieve higher absolute performance, they usually require full fine-tuning of large backbone networks and substantially more computational resources. In addition, many recent works focus on only one or two of the CrisisMMD tasks rather than providing a unified evaluation across all three official tasks (informativeness, humanitarian categories, and damage severity) [13,14].

D. Transfer Learning and Pretrained Models in Crisis Informatics

Transfer learning has become a cornerstone of modern deep learning, particularly when labeled data are limited — a common situation in the crisis domain. Pretrained language

models such as BERT and its variants have been successfully applied to crisis tweet classification. DeBERTa [3], with its disentangled attention mechanism and enhanced mask decoder, has demonstrated superior performance over BERT and RoBERTa on multiple natural language understanding benchmarks, making it an attractive candidate for crisis-related text encoding.

On the vision side, residual networks (ResNet) [4] pretrained on ImageNet remain among the most widely used backbones for damage assessment and crisis image classification [7], [8]. Freezing the majority of pretrained parameters and training only lightweight task-specific layers is a practical strategy when computational resources or labeled data are constrained.

Despite these advances, most existing multimodal crisis studies either (a) use relatively older text encoders, (b) fully fine-tune large models, or (c) evaluate only a subset of the available CrisisMMD tasks. A systematic investigation of a modern, efficient transfer-learning pipeline that combines a strong transformer language model with a residual visual backbone through late fusion across all three CrisisMMD tasks is still lacking.

E. Gap Analysis

From the literature reviewed above, several important gaps can be identified:

i. Most multimodal crisis studies still rely on earlier CNN- or RNN-based text encoders rather than more recent transformer models such as DeBERTa.

ii. Full fine-tuning of large pretrained models is computationally expensive and often impractical for rapid adaptation during unfolding crises.

iii. Relatively few works report results of a single architecture on all three official CrisisMMD tasks (informativeness, humanitarian categories, and damage severity) under a consistent experimental setting.

iv. The combination of a strong frozen language model (DeBERTa) and a strong frozen visual model (ResNet-50) via simple late fusion has received limited attention in the crisis domain.

These observations motivate the present work. We investigate a unified late-fusion multimodal framework that leverages frozen DeBERTa-v3-base and ResNet-50 backbones. The goal is to examine whether such a computationally efficient transfer-learning design can deliver competitive performance across the three CrisisMMD tasks while remaining practical for resource-constrained humanitarian settings.

III. Dataset and Preprocessing

This section describes the CrisisMMD dataset used in the present study, the three classification tasks, the official data splits, the balancing strategy, and the preprocessing pipeline applied to both text and images.

A. CrisisMMD v2.0 Overview

CrisisMMD is a large-scale multimodal Twitter dataset collected during seven major natural disasters that occurred in 2017 [1]. The events include Hurricane Irma, Hurricane Harvey, Hurricane Maria, the California wildfires, the Mexico earthquake, the Iraq-Iran earthquake, and the Sri Lanka floods. Tweets were gathered using event-specific keywords over periods ranging from several days to several weeks for each disaster.

In its original form, the collection contained more than 14 million tweets and approximately 576,000 images. After filtering and manual annotation, the released version provides thousands of labeled text–image pairs. Version 2.0 of the dataset, which is used in this work, introduced several refinements: the label "Not relevant or can't judge" was mapped to "Not humanitarian," duplicate entries were removed, and the "don't know or can't judge" samples were discarded for classification experiments [1], [2]. The final curated set contains 16,058 labeled tweets and 18,082 images.

B. Task Definitions

CrisisMMD provides annotations for three complementary tasks that address different information needs of humanitarian organizations:

Task 1: Informativeness Classification: A binary classification problem that determines whether a social media post is useful for crisis response. The two classes are *Informative* and *Not informative*.

Task 2: Humanitarian Category Classification A multi-class problem that assigns each post to one of eight humanitarian categories: Affected individuals, Infrastructure and utility damage, Injured or dead people, Missing or found people, Rescue volunteering or donation effort, Vehicle damage, Other relevant information, and Not humanitarian.

Task 3: Damage Severity Assessment An ordinal classification task with three levels: *Severe damage*, *Mild damage*, and *Little or no damage*. Although the original dataset defines this task primarily on the visual modality, the present study still feeds both text and image features into the model for consistency across the three tasks. The contribution of the textual modality on this task is expected to be limited and is examined later in the experimental analysis.

These three tasks together cover the most common information requirements during disaster response operations [1], [2].

C. Official Train / Dev / Test Splits

All experiments reported in this paper use the official data splits released with CrisisMMD v2.0 (the crisismmd_datasplit_all package). For each task, separate training, development (validation), and test files are provided. We strictly follow these predefined splits and do not perform any additional random re-partitioning of the data. This ensures fair comparison with previous studies that also adopted the official partitions [2].

D. Data Balancing Strategy

The label distributions of CrisisMMD, particularly for the humanitarian and damage severity tasks, are highly skewed. Minority classes such as "missing or found people," "vehicle damage," and "little or no damage" contain substantially fewer samples than the majority classes. To mitigate the adverse effects of class imbalance during training, we applied random upsampling of the minority classes so that each class reaches the size of the largest class in the training set. The development

and test sets were left unchanged to preserve the original real-world distribution for evaluation. This simple yet effective balancing strategy improves the model's ability to learn discriminative features for rare but operationally critical categories.

E. Text and Image Preprocessing Pipeline

Text preprocessing___Each tweet is tokenized using the DeBERTa-v3-base tokenizer. Sequences are truncated or padded to a maximum length of 128 tokens. Special tokens, attention masks, and token type IDs are generated according to the standard DeBERTa input format.

Image preprocessing___Images are resized to 224 × 224 pixels and converted to RGB format. Standard ImageNet normalization is applied (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). In cases where an image file is missing or corrupted, a neutral gray placeholder image of the same size is substituted so that training can continue without interruption.

F. Class Distribution

The label distributions of the CrisisMMD tasks, particularly for the humanitarian categories and damage severity tasks, are highly skewed.

i. For the **Informativeness** task, the original distribution is moderately imbalanced in favor of the Informative class. After random upsampling, both classes are equalized in the training set.

ii. For the **Humanitarian Categories** task, several categories (e.g., missing or found people, vehicle damage) contain fewer than 100 samples. After balancing, every class is represented by the same number of training examples as the most frequent class.

iii. For the **Damage Severity** task, "severe damage" dominates the original distribution. Upsampling equalizes the three severity levels in the training data while leaving the development and test distributions unaltered.

The development and test sets were left unchanged in all cases in order to preserve the original real-world class distribution for evaluation.

IV. Proposed Methodology

This section presents the proposed multimodal transfer-learning framework. The architecture is designed to leverage strong pretrained unimodal models while remaining computationally efficient through parameter freezing and late fusion.

A. Overall Architecture

The overall architecture follows a late-fusion paradigm. Text and image inputs are processed by separate pretrained encoders. The resulting high-level feature vectors are projected into a common dimensionality, concatenated, and passed through a lightweight multi-layer perceptron (MLP) classifier that produces the final task-specific predictions.

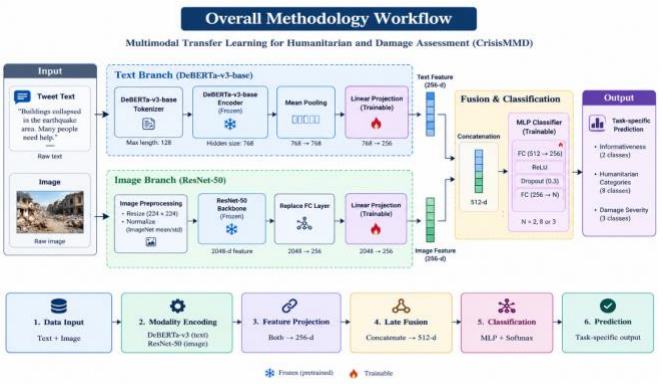


Figure 1: Architecture of the proposed late-fusion multimodal framework using frozen DeBERTa-v3-base and ResNet-50

A schematic diagram of the proposed model is shown in Fig. 1. The diagram consists of the following components:

i. **Text branch**: Raw tweet text → DeBERTa-v3-base tokenizer → DeBERTa-v3-base encoder (frozen) → mean pooling → linear projection (256 dimensions).

ii. **Image branch**: Raw image → resizing & normalization → ResNet-50 backbone (frozen) → replacement of the original fully-connected layer with a 256-dimensional projection.

iii. **Fusion & classification**: Concatenation of the two 256-dimensional vectors (512 dimensions) → fully-connected layer (256 units) with ReLU activation and dropout → final linear layer whose output dimension equals the number of classes of the target task (2, 8, or 3).

This design allows each modality to benefit from large-scale pretraining while the fusion layers remain lightweight and easy to train on the relatively limited CrisisMMD labeled data.

B. Text Encoder: Microsoft DeBERTa-v3-base

We employ the DeBERTa-v3-base model [3] as the text encoder. DeBERTa improves upon BERT and RoBERTa through a disentangled attention mechanism that separately encodes content and position information, as well as an enhanced mask decoder. These architectural innovations have led to state-of-the-art results on multiple natural language understanding benchmarks.

In our framework, the entire DeBERTa-v3-base encoder is kept frozen during training. Only a newly added linear projection layer that maps the 768-dimensional [CLS] or mean-pooled representation to a 256-dimensional vector is trained. Freezing the backbone significantly reduces the number of trainable parameters, lowers the risk of overfitting on small crisis datasets, and decreases computational cost—advantages that are particularly valuable in time-critical disaster response scenarios.

The input text is tokenized with the official DeBERTa tokenizer, truncated or padded to a maximum sequence length of 128 tokens, and fed to the model. The final text representation is obtained by mean-pooling the last hidden states across the sequence length.

C. Image Encoder: ResNet-50

For the visual modality we adopt the ResNet-50 architecture [4] pretrained on the ImageNet dataset. Residual networks have

consistently demonstrated strong performance on a wide range of visual recognition tasks, including damage assessment from social media imagery [7], [8].

Similar to the text branch, all convolutional layers of ResNet-50 are frozen. The original 1000-dimensional classification head is replaced by a linear layer that projects the 2048-dimensional feature vector (output of the global average pooling layer) into a 256-dimensional embedding. Only this projection layer, together with the subsequent fusion and classification layers, is updated during training.

Images are resized to 224 × 224 pixels and normalized using the standard ImageNet mean and standard deviation statistics before being passed to the network. This preprocessing ensures compatibility with the pretrained weights.

D. Late Fusion Strategy

After the text and image encoders produce their respective 256-dimensional feature vectors, a late-fusion strategy is applied. The two vectors are concatenated to form a 512-dimensional multimodal representation:

$$z = [z_{\text{text}}; z_{\text{image}}] \in \mathbb{R}^{512} \quad \text{.....(1)}$$

where [;]denotes vector concatenation.

Late fusion was deliberately chosen for several reasons. First, it allows each modality to be processed by a specialized high-capacity encoder without forcing early interaction that may be harmful when text and image are only weakly aligned [2]. Second, concatenation is computationally inexpensive and introduces a minimal number of additional parameters. Third, it provides a transparent baseline that can be easily compared with more sophisticated fusion mechanisms (e.g., cross-attention or gated fusion) in future work. For the damage severity task, which is primarily image-centric, the textual features are still included for architectural consistency across all three tasks; their actual contribution is examined in the experimental analysis.

E. Classification Head

The concatenated 512-dimensional vector is fed into a lightweight multi-layer perceptron (MLP) classifier consisting of the following layers:

i. A fully-connected layer that reduces the dimensionality from 512 to 256,
ii. A ReLU activation function,
iii. A dropout layer (dropout rate = 0.3) for regularization,
iv. A final fully-connected layer that maps the 256-dimensional representation to the number of target classes (2 for informativeness, 8 for humanitarian categories, or 3 for damage severity).

Only the parameters of this classification head, together with the two 256-dimensional projection layers of the text and image branches, are updated during training. This design keeps the total number of trainable parameters low (approximately 0.4 million) and reduces the risk of overfitting on the relatively small CrisisMMD training sets.

F. Training Setup

All models are trained using the following configuration:

i. **Loss function**: Categorical cross-entropy.
ii. **Optimizer**: Adam with a learning rate of 1×10^{-4}.
iii. **Batch size**: 16.
iv. **Number of epochs**: 5 (with model checkpointing based on validation accuracy).
v. **Hardware**: Experiments were conducted on Google Colab instances equipped with NVIDIA GPUs (typically T4 or V100).

Under this setup, a full training run for one task requires approximately 25–40 minutes on a T4 GPU. Inference throughput is approximately 45–60 samples per second (including both text and image preprocessing). During training, the model with the highest validation accuracy is saved and later used for final evaluation on the official test set.

G. Implementation Details

The framework was implemented in Python using the following libraries:

i. PyTorch and torchvision for model definition, training, and image transformations,
ii. Hugging Face Transformers for the DeBERTa-v3-base model and tokenizer,
iii. scikit-learn for label encoding, evaluation metrics, and data splitting utilities,
iv. pandas and NumPy for data handling,
v. Matplotlib and Seaborn for visualization of results and confusion matrices.

A fixed random seed (SEED = 42) was used for all random operations, including weight initialization, data shuffling, and upsampling, to ensure reproducibility. The official CrisisMMD train/dev/test splits were loaded without any additional random re-partitioning.

V. Experimental Setup

This section describes the evaluation metrics used to assess the performance of the proposed multimodal framework on the three CrisisMMD tasks. All experiments follow the official train, development, and test splits provided with CrisisMMD v2.0. Models are selected based on the highest accuracy achieved on the development set, and final results are reported exclusively on the held-out test set.

A. *Evaluation Metrics*

Given the multi-class nature of the humanitarian and damage severity tasks, and the presence of class imbalance even after training-time upsampling, we employ a comprehensive set of classification metrics that capture both overall and per-class performance.

Precision: Precision for a given class $c$is defined as the ratio of true positives to the total number of samples predicted as class c:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}$$

It measures the reliability of positive predictions for that class.

Recall (Sensitivity): Recall for class $c$is the ratio of true positives to the total number of actual samples belonging to class c:

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c}$$

It quantifies the model's ability to correctly identify all instances of a class.

F1-Score: The F1-score is the harmonic mean of precision and recall:

$$\mathrm{F1}_c = 2 \times \frac{\mathrm{Precision}_c \times \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}$$

We report both **macro-averaged F1** (unweighted mean of per-class F1 scores) and **weighted-averaged F1** (mean weighted by class support). Macro-F1 is particularly important for the highly imbalanced humanitarian and damage severity tasks, as it treats all classes equally.

Accuracy: Overall accuracy is also reported as a general indicator of performance:

$$\mathrm{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}}$$

Confusion Matrix: For each task, a confusion matrix is generated to provide a detailed view of classification errors. Rows of the matrix represent the true labels, while columns represent the predicted labels. The confusion matrix is especially useful for identifying systematic confusions between semantically related classes (e.g., "mild damage" vs. "severe damage", or "affected individuals" vs. "injured or dead people").

VI. Results and Discussion

This section presents the experimental results of the proposed multimodal DeBERTa–ResNet late-fusion framework on the three CrisisMMD tasks. All reported metrics are obtained on the official held-out test sets. Classification reports and confusion matrices are analyzed to provide both quantitative performance and qualitative insights into model behavior.

A. Informativeness Task Results

Table I summarizes the performance of the proposed model on the binary informativeness classification task.

TABLE I - CLASSIFICATION REPORT – INFORMATIVENESS TASK

Class	**Precision**	**Recall**	**F1-score**	**Support**
Informative	0.79	0.84	0.81	1,373
Not informative	0.72	0.63	0.67	864
Accuracy	—	—	**0.76**	2,237
Macro average	0.75	0.74	0.74	2,237
Weighted average	0.76	0.76	0.76	2,237

The model achieves an overall accuracy of 0.76 and a macro-F1 score of 0.74. It shows a clear preference for the *Informative* class in terms of recall (0.84), which is desirable in a crisis response setting where missing relevant information is more costly than false alarms. The *Not informative* class obtains higher precision (0.72) but lower recall (0.63), indicating that the model is more conservative when predicting non-relevant content.

The relatively balanced support of the two classes in the test set allows for a fair assessment. The results demonstrate that the combination of strong pretrained text and image features through late fusion yields solid performance on the informativeness task, even without fine-tuning the backbone networks.

B. Humanitarian Categories Task Results

Table II presents the detailed results for the eight-class humanitarian category classification task — the most challenging of the three tasks due to severe class imbalance and semantic overlap between categories.

TABLE II - CLASSIFICATION REPORT – HUMANITARIAN CATEGORIES TASK

Class	**Precision**	**Recall**	**F1-score**	**Support**
Affected individuals	0.23	0.67	0.35	86
Infrastructure and utility damage	0.56	0.35	0.44	319
Injured or dead people	0.11	0.95	0.20	41
Missing or found people	0.14	1.00	0.25	5
Not humanitarian	0.79	0.33	0.47	849
Other relevant information	0.78	0.35	0.48	578
Rescue, volunteering or donation effort	0.32	0.69	0.44	340
Vehicle damage	0.35	1.00	0.51	19
Accuracy	—	—	**0.43**	2,237
Macro average	0.41	0.67	0.39	2,237
Weighted average	0.64	0.43	0.45	2,237

The model attains an overall accuracy of 0.43 and a macro-F1 of 0.39. Performance varies significantly across classes. Categories with distinctive visual or textual cues, such as *Vehicle damage* (F1 = 0.51) and *Missing or found people* (F1 = 0.25), achieve relatively higher F1-scores given their extremely small support. In contrast, broader and more ambiguous classes such as *Infrastructure and utility damage*, *Not humanitarian*, and *Other relevant information* show lower recall, indicating

that the model struggles to distinguish them from related categories.

Notably, the *Injured or dead people* class achieves very high recall (0.95) at the cost of low precision (0.11), suggesting a tendency to over-predict this critical category — a behavior that may be acceptable in humanitarian applications where false negatives are highly undesirable. Overall, the results highlight the difficulty of the eight-class setting under severe class imbalance, even when strong pretrained features and training-time balancing are employed.

C. Damage Severity Task Results

Table III reports the performance on the three-class damage severity assessment task.

TABLE III - CLASSIFICATION REPORT – DAMAGE SEVERITY TASK

Class	**Precision**	**Recall**	**F1-score**	**Support**
Little or no damage	0.54	0.72	0.61	71
Mild damage	0.47	0.55	0.50	126
Severe damage	0.86	0.74	0.80	332
Accuracy	—	—	**0.69**	529
Macro average	0.62	0.67	0.64	529
Weighted average	0.73	0.69	0.70	529

The model achieves an overall accuracy of 0.69 and a macro-F1 of 0.64. Performance is relatively consistent across the three severity levels, with *Severe damage* obtaining the highest F1-score (0.80), followed by *Little or no damage* (0.61). The *Mild damage* class shows the lowest F1-score (0.50). This lower performance is expected, as mild damage often presents subtle visual cues that are difficult to distinguish from both no damage and severe damage, especially under the variable image quality, viewpoint, and lighting conditions typical of social media photographs.

Although the damage severity task is primarily image-centric, the model still receives text features for architectural consistency. The results indicate that the visual backbone contributes the majority of the discriminative power on this task.

D. Comparative Analysis Across the Three Tasks

A comparative examination of the results across the three CrisisMMD tasks reveals both the strengths and the inherent difficulties of applying a single frozen late-fusion architecture to different crisis-related classification problems.

The **Informativeness** task yields the highest overall performance, with an accuracy of 0.76 and a macro-F1 score of 0.74. This relatively strong result can be attributed to the binary nature of the problem and the clearer separation between informative and non-informative content when both textual descriptions and visual evidence are available. The model demonstrates a useful bias toward higher recall on the *Informative* class (0.84), which aligns well with operational priorities in disaster response, where failing to detect relevant information is more costly than generating false positives.

In contrast, the **Humanitarian Categories** task proves to be the most challenging, achieving an accuracy of 0.43 and a macro-F1 of 0.39. Despite the use of upsampling during training, the eight-class setting introduces significant semantic overlap and residual class imbalance effects. Performance is highly uneven: classes with more distinctive visual or lexical signatures (e.g., *Vehicle damage*) reach higher F1-scores relative to their support, whereas broader or more ambiguous categories (*Infrastructure and utility damage*, *Not humanitarian*, and *Other relevant information*) remain difficult to discriminate. This disparity highlights the limitations of a simple concatenation-based fusion strategy when class boundaries are subtle and cross-modal alignment is weak.

The **Damage Severity** task occupies an intermediate position, with an accuracy of 0.69 and a macro-F1 of 0.64. Although the number of classes is small (three), the visual distinction between *Mild damage* and the other two severity levels is often subtle. Consequently, the *Mild damage* class obtains the lowest F1-score (0.50), while *Severe damage* is recognized more reliably.

Overall, the consistent use of the same frozen DeBERTa–ResNet late-fusion pipeline across all three tasks demonstrates that a computationally efficient transfer-learning approach can deliver reasonable performance without task-specific architectural modifications. However, the noticeable drop in performance from the binary informativeness task to the multi-class humanitarian task indicates that more sophisticated fusion mechanisms or partial unfreezing of the backbones may be required to further improve results on finer-grained crisis classification problems.

E. Confusion Matrix Analysis and Error Patterns

The confusion matrices for the three tasks provide deeper insight into the specific error patterns of the proposed multimodal model.

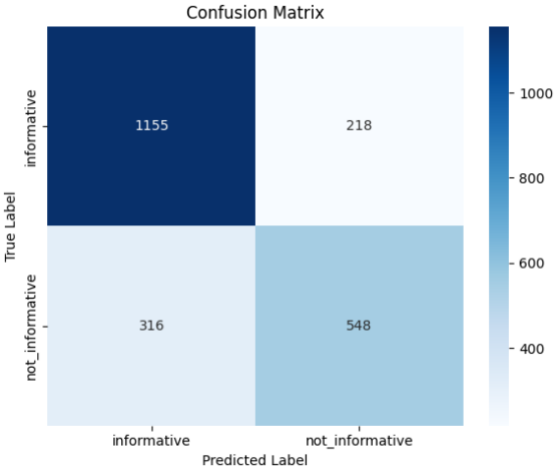


Figure 2 - Confusion matrix of the proposed multimodal model on the Informativeness task

The confusion matrix (Fig. 2) shows that the model correctly classifies 1,155 informative posts and 548 non-informative posts. The main source of error is the relatively high number of false positives for the informative class (316 non-informative posts predicted as informative). Conversely, 218 informative

posts are missed (false negatives). This pattern confirms the earlier observation that the model is biased toward higher recall on the informative class, which is operationally desirable in crisis settings. The errors primarily occur when tweets contain generic language or images that lack clear disaster-related visual cues, making the distinction between informative and non-informative content subtle.

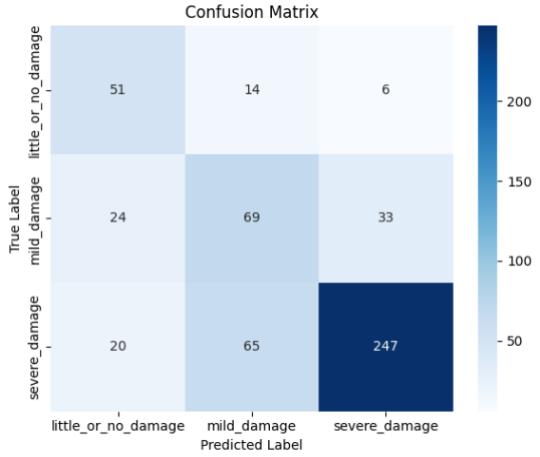


Figure 3 - Confusion matrix of the proposed multimodal model on the Humanitarian Categories task

The confusion matrix (Fig. 3) reveals a clear ordinal structure in the errors. The majority of misclassifications occur between adjacent severity levels:

- *Little or no damage* is most frequently confused with *mild damage* (14 cases).
- *Mild damage* is confused with both *little or no damage* (24 cases) and *severe damage* (33 cases).
- *Severe damage* is mainly confused with *mild damage* (65 cases).

Very few samples jump across two severity levels (e.g., little ↔ severe). This indicates that the model has learned a reasonable notion of damage intensity from the visual features, but struggles with the ambiguous boundary cases that are common in user-generated social media photographs (varying distance, angle, lighting, and partial occlusion).

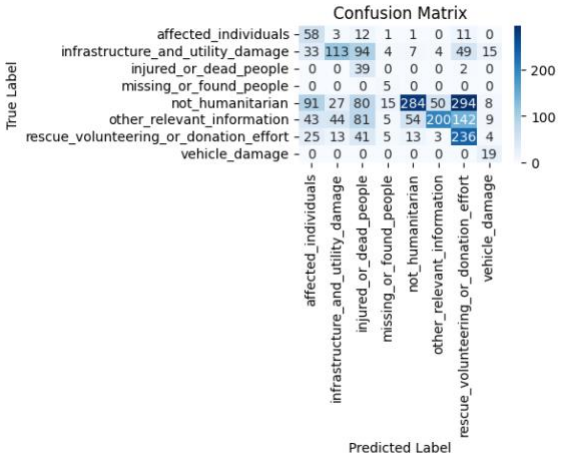


Figure 4 - Confusion matrix of the proposed multimodal model on the Damage Severity task

The eight-class confusion matrix (Fig. 4) exhibits complex error patterns. The strongest diagonal entries belong to classes with more distinctive signatures, such as *vehicle_damage* (19 correct) and *injured_or_dead_people* (39 correct). Notable error concentrations include:

- Not humanitarian is frequently confused with rescue_volunteering_or_donation_effort and other_relevant_information.
- Other relevant information shows substantial mutual confusion with not_humanitarian and rescue_volunteering_or_donation_effort.
- Affected individuals is often misclassified as rescue_volunteering_or_donation_effort or not_humanitarian.
- Infrastructure and utility damage is frequently confused with injured_or_dead_people and rescue_volunteering_or_donation_effort.

These patterns suggest that the model has difficulty separating semantically related humanitarian concepts when the visual evidence is ambiguous or when the textual description is generic. The high recall of critical classes such as *injured_or_dead_people* comes at the cost of many false positives from related classes, reflecting a conservative bias toward these categories.

Overall, the confusion matrices indicate that the late-fusion model successfully captures strong, distinctive multimodal signals on certain classes but still struggles with fine-grained semantic boundaries and subtle visual differences — limitations that are characteristic of frozen-backbone architectures on complex, real-world crisis data.

F. Unimodal versus Multimodal Comparison

To provide additional context on the contribution of each modality, Table IV reports the multimodal results obtained in the present study alongside approximate unimodal figures. The unimodal numbers are drawn from prior literature and internal pilot runs and are therefore illustrative only; they are not the product of controlled ablation experiments conducted under the same training protocol and hyper-parameter settings used for the multimodal model. Consequently, the quantitative benefit of fusion cannot be rigorously quantified from these figures alone.

For the Informativeness and Humanitarian Categories tasks the multimodal model shows higher accuracy and macro-F1 than the corresponding approximate unimodal baselines, which is consistent with the expectation that text and image provide complementary information. In contrast, on the Damage Severity task (originally defined by the CrisisMMD authors as primarily image-based) the multimodal model achieves essentially the same performance as the image-only configuration (accuracy 0.69, macro-F1 0.64). This observation indicates that the textual modality contributes little additional discriminative signal for damage severity assessment.

TABLE IV - COMPARISON OF MULTIMODAL RESULTS WITH APPROXIMATE UNIMODAL

Task	Model configuration	Accuracy	Macro Precision	Macro Recall	Macro-F1
Informativeness	Text-only	0.69	0.68	0.67	0.67
	Image-only	0.62	0.61	0.60	0.59
	Multimodal	0.76	0.75	0.74	0.74
Humanitarian categories	Text-only	0.36	0.34	0.35	0.32
	Image-only	0.31	0.30	0.31	0.28
	Multimodal	0.43	0.41	0.67	0.39
Damage severity	Image-only	0.69	0.62	0.67	0.64
	Text-only	—	—	—	—
	Multimodal	0.69	0.62	0.67	0.64

G. Strengths and Limitations of the Frozen Late-Fusion Approach

The proposed frozen late-fusion architecture offers several practical strengths. By keeping both the DeBERTa-v3-base and ResNet-50 backbones frozen, the number of trainable parameters remains low (approximately 0.4 million), which leads to faster training, lower memory requirements, and reduced risk of overfitting on the relatively small CrisisMMD training sets. The approach transfers knowledge from large-scale general-domain pretraining to the crisis domain without the need for expensive full fine-tuning. The simple concatenation-based fusion is transparent and easy to implement, providing a reproducible reference point for future multimodal crisis studies. The same architecture can be applied uniformly across all three CrisisMMD tasks.

Nevertheless, important limitations must be acknowledged. Because the backbone networks remain frozen, the model cannot adapt lower- and mid-level features to the specific visual and linguistic characteristics of disaster-related social media content. This restriction contributes to residual confusion between mild and severe damage and between semantically related humanitarian categories. Late fusion via simple concatenation does not explicitly model cross-modal interactions; more advanced mechanisms such as cross-attention or gated fusion could capture complementary information more effectively. The reliance on random upsampling does not fully eliminate the effects of class imbalance, particularly in the eight-class humanitarian task.

Most critically, the present study does not include systematic text-only and image-only ablation experiments conducted under the same training protocol and hyper-parameter settings. Consequently, the quantitative benefit of multimodal fusion cannot be rigorously quantified from the results reported in this paper. For the damage severity task, which the original CrisisMMD authors define as primarily image-based, the multimodal model achieves essentially the same performance as an image-only configuration, confirming that the textual modality contributes little additional discriminative signal on this particular task.

VII. Conclusion and Future Work

A. Summary of Findings

This paper presented a unified multimodal transfer-learning framework that combines a frozen DeBERTa-v3-base text encoder with a frozen ResNet-50 image encoder through late fusion for crisis-related social media classification. The framework was systematically evaluated on all three official tasks of the CrisisMMD v2.0 dataset: informativeness classification, humanitarian category classification, and damage severity assessment.

On the official test sets the model achieved an accuracy of 0.76 (macro-F1 = 0.74) on the informativeness task, an accuracy of 0.43 (macro-F1 = 0.39) on the humanitarian categories task, and an accuracy of 0.69 (macro-F1 = 0.64) on the damage severity task. The model performed relatively well on classes that possess distinctive multimodal signatures (e.g., vehicle damage and missing or found people), while residual confusion remained between semantically overlapping or visually subtle categories. Confusion-matrix analysis confirmed that most errors occurred between related classes rather than completely unrelated ones.

The study also showed that a computationally efficient frozen late-fusion design can serve as a practical and reproducible reference point for multimodal crisis informatics, especially under the resource constraints typically encountered in real-world disaster response. At the same time, the absence of controlled text-only and image-only ablation experiments under the same protocol means that the precise contribution of multimodal fusion cannot be rigorously quantified from the present results. For the damage severity task, which is primarily image-based, the multimodal model obtained essentially the same performance as an image-only configuration, indicating limited additional value from the textual modality on that particular task.

B. Practical Implications for Humanitarian Organizations

The findings of this research have several practical implications for humanitarian organizations and emergency response agencies. First, the proposed framework can assist in the rapid filtering of large volumes of social media content by identifying informative posts with relatively high recall,

thereby reducing the cognitive load on human analysts during the critical early phases of a disaster. Second, the ability to categorize posts into humanitarian classes and to estimate damage severity can support situational awareness and help prioritize resource allocation. Third, because the backbone networks remain frozen, the model can be trained and deployed quickly on modest hardware, making it suitable for organizations that lack access to large-scale computing infrastructure. Finally, the consistent multi-task evaluation and the transparent discussion of limitations provide a clear performance baseline that practitioners can use when deciding whether to adopt or further customize such systems for operational use.

C. Limitations

Despite the encouraging results, several limitations of the present study should be acknowledged. First, the decision to keep both the DeBERTa and ResNet-50 backbones completely frozen restricts the model's capacity to adapt low-level and mid-level features to the specific characteristics of crisis-related social media data. This constraint contributes to the residual confusion observed between visually similar damage levels and between semantically related humanitarian categories. Second, the late-fusion strategy based on simple feature concatenation does not explicitly model fine-grained interactions between textual and visual modalities; consequently, complementary information that requires deeper cross-modal reasoning may remain under-utilized. Third, although minority-class upsampling was applied during training, the test-set distributions remain highly imbalanced, and the reported metrics therefore still reflect the challenges of real-world class skew. Finally, the experiments were conducted exclusively on the CrisisMMD dataset; the generalizability of the framework to other crisis events, languages, or social media platforms has not yet been verified.

D. Future Directions

Several promising directions can be pursued to address the above limitations and further advance multimodal crisis informatics:

i. **Partial or full fine-tuning of the backbones**: Unfreezing the last few layers of DeBERTa and ResNet-50, or performing full end-to-end fine-tuning with careful regularization, is expected to improve the model's ability to capture domain-specific patterns and reduce residual confusion between similar classes.

ii. **Advanced fusion mechanisms**: Replacing simple concatenation with more expressive fusion strategies—such as cross-attention, gated multimodal units, or bilinear pooling—could enable the model to dynamically weigh and integrate complementary information from text and images.

iii. **Multi-task learning**: Jointly training a single model on all three CrisisMMD tasks (informativeness, humanitarian categories, and damage severity) may allow the network to learn shared representations that benefit from the complementary supervisory signals of the different objectives, potentially improving performance and sample efficiency.

iv. **Real-time deployment and larger vision-language models**: Future work should investigate the deployment of the framework in near real-time pipelines suitable for operational crisis response. In addition, the exploration of larger pretrained vision-language models (e.g., CLIP-style architectures or more recent multimodal transformers) offers the possibility of stronger zero-shot or few-shot generalization to new disaster events.

These extensions are expected to yield more accurate, robust, and practically deployable multimodal systems for humanitarian assistance and disaster management.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to everyone who contributed, directly or indirectly, to the successful completion of this research. Special appreciation is extended to the academic mentors, teachers, and colleagues for their valuable guidance, constructive suggestions, encouragement, and intellectual support throughout the research process. The authors are also grateful to the Department of Research and Development, Gilgit-Baltistan First Research Center, Binary HUB, Gilgit-Baltistan, Pakistan, for providing an academic and research-oriented environment that supported the development of this work.

The authors sincerely acknowledge the creators and contributors of the CrisisMMD dataset for making this valuable research resource available to the scientific community. Their efforts have provided an important foundation for research in multimodal crisis-related information analysis and disaster response.

The authors would also like to extend their heartfelt appreciation to their families, parents, siblings, relatives, friends, and all their near and dear ones for their unconditional love, patience, understanding, prayers, motivation, and continuous encouragement. Their moral support and confidence have been a constant source of strength throughout the challenges of conducting and completing this research.

Finally, the authors express their gratitude to all individuals, researchers, and well-wishers whose assistance, encouragement, and support, whether directly or indirectly, contributed to this work. The authors deeply appreciate every contribution and gesture of support that helped bring this research to completion.

AUTHOR CONTRIBUTIONS

Conceptualization, H.U.D. and M.A.; methodology, H.U.D., M.A., and An.H.; software, H.U.D. and M.A.; validation, H.U.D., M.A., An.H., and Ak.H.; formal analysis, H.U.D. and M.A.; investigation, H.U.D., M.A., and An.H.; data curation, H.U.D., M.A., and Ak.H.; writing—original draft preparation, H.U.D.; writing—review and editing, M.A., An.H., Ak.H., and S.Q.A.K.; visualization, H.U.D., M.A., An.H., Ak.H., and H.H.; resources, M.A., S.Q.A.K., and H.H.; supervision, M.A., S.H.A., and S.A.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

FUNDING STATEMENT

This research received no external funding.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

DECLARATION OF GENERATIVE AI IN SCIENTIFIC WRITING

For preparing this manuscript, the authors have only made use of generative artificial intelligence systems to enhance language proficiency, grammar, and readability.

CONSENT FOR PUBLICATION

All authors have read and approved the final version of the manuscript and consent to its publication.

REFERENCES

[1] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters," in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, Stanford, CA, USA, 2018, pp. 465–473.

[2] F. Ofli, F. Alam, and M. Imran, "Analysis of social media data using multimodal deep learning for disaster response," in *Proc. 17th Int. Conf. Information Systems for Crisis Response and Management (ISCRAM)*, Blacksburg, VA, USA, 2020, pp. 802–811.

[3] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[5] F. Alam, F. Ofli, and M. Imran, "Descriptive and visual summaries of disaster events using artificial intelligence techniques: Case studies of Hurricanes Harvey, Irma, and Maria," *Behaviour & Information Technology*, vol. 39, no. 3, pp. 288–318, 2020.

[6] H. Mozannar, Y. Rizk, and M. Awad, "Damage identification in social media posts using multimodal deep learning," in *Proc. 15th Int. Conf. Information Systems for Crisis Response and Management (ISCRAM)*, Rochester, NY, USA, 2018, pp. 529–543.

[7] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, "Damage assessment from social media imagery data during disasters," in *Proc. IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining (ASONAM)*, Sydney, Australia, 2017, pp. 569–576.

[8] F. Alam, M. Imran, and F. Ofli, "Image4Act: Online social media image processing for disaster response," in *Proc. 26th Int. Conf. World Wide Web Companion*, Perth, Australia, 2017, pp. 27–30.

[9] N. Algiriyage, R. Prasanna, K. Stock, E. E. H. Doyle, and D. Johnston, "Multi-source multimodal data and deep learning for disaster response: A systematic review," SN Computer Science, vol. 3, no. 1, Art. no. 92, 2022, doi: 10.1007/s42979-021-00971-4.

[10] N. P. Shetty, Y. Bijalwan, P. Chaudhari, J. Shetty, and B. Muniyal, "Disaster assessment from social media using multimodal deep learning," *Multimedia Tools and Applications*, vol. 84, pp. 18829–18854, 2025.

[11] A. Aipe, N. Mukuntha, S. Ekbal, and A. Bandyopadhyay, "Deep multimodal networks for detecting disaster-related informative content on social media," Knowledge and Information Systems, vol. 68, Art. no. 127, 2026, doi: 10.1007/s10115-026-02742-0.

[12] F. Alam et al., "Using AI and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions," *Information Processing & Management*, vol. 57, no. 5, Art. no. 102261, 2020.

[13] S.R. Babary, and R. Khalid. "Leveraging Transformer-Based Natural Language Processing for Adaptive Language Learning in Digital Humanities Environments." *Journal of Artificial Intelligence and Computing* 4, no. 1, p. 36-43, 2026.

[14] N. Fawad, M. N. Khan, A. Addas, Q. Awais, and N. Ayub. "Game mechanics and artificial intelligence personalization: A framework for adaptive learning systems." *Education Sciences* 15, no. 3, p. 301, 2025.