IDRiM (2026) 16 (1) **ISSN: 2185-8322**
DOI10.5595/001c.162527

Journal of Integrated Disaster Risk Management

Regular Article

Cascaded Attention-Based Multimodal Framework for Robust Disaster Tweet Classification

Ankita Dixit[1*] and Indu Chawla[1]

Received: 12/08/2025 / Accepted: 12/02/2026 / Published online: 10/06/2026

Abstract Disasters require timely and informed responses, and social media platforms like Twitter (now X) have become vital sources of real-time information. Unimodal methods rely on a single data source that leverages either text or visual modalities, resulting in incomplete situational awareness. However, tweets often contain noisy, brief, and embedded image information, making classification a challenging task. To effectively harness textual and visual modalities, we propose a cascaded attention-based multimodal framework that improves the precision and relevance of disaster-related insights. The model employs a dual-stream architecture that separately encodes textual and visual features, followed by a cascaded attention mechanism that dynamically refines and fuses modality-specific information. Our approach leverages both intra-modal attention to highlight informative features within each modality and inter-modal attention to align and fuse complementary semantics across text and visuals. Experiments conducted on the benchmark disaster tweet dataset demonstrate that the proposed framework significantly outperforms existing unimodal and multimodal baselines in terms of accuracy and robustness. The proposed framework primarily contributes to research in multimodal disaster analysis. It can be integrated into social media monitoring and decision-support systems used by humanitarian organizations, emergency management agencies, and public safety authorities to enhance situational awareness and support faster, more informed crisis response.

Keywords: social media, disaster, unimodal, multimodal, image, text, self-attention, cross attention

[1] Department of Computer Science & Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India
* Corresponding author email: ankitadixit177@gmail.com

1. INTRODUCTION

The proliferation of social media platforms has transformed disaster response by providing real-time, location-specific information from eyewitnesses. Platforms such as Twitter enable users to report injuries, fatalities, and infrastructure damage through text, images, and videos (Acerbo et al., 2017; Seneviratne et al.,2024). Early disaster tweet analysis systems predominantly relied on unimodal approaches, processing either textual or visual content independently due to their lower computational complexity and ease of deployment. Text-only models have been widely adopted because textual information is more consistently available than images during emergencies. However, unimodal methods often provide incomplete situational understanding, as disaster-related tweets frequently contain complementary information across multiple modalities. Disaster-related tweets can broadly be categorized as informative or non-informative. Informative tweets contain actionable or situationally relevant information useful for response and relief efforts, such as reports of casualties, infrastructure damage, or requests for assistance, whereas non-informative tweets include expressions of emotion, opinions, or general commentary unrelated to emergency operations. Humanitarian organizations increasingly rely on the timely identification of informative content to enhance situational awareness and optimize resource allocation during crisis events (Devaraj et al., 2020).

However, unimodal methods often provide incomplete situational understanding, as disaster-related tweets frequently contain complementary information across multiple modalities. This limitation has motivated recent research into multimodal approaches that jointly exploit textual and visual cues to capture richer semantic context. Nevertheless, the high volume, noise, and heterogeneity of multimodal crisis data pose significant challenges for manual analysis and model design, necessitating robust automated frameworks capable of effectively fusing multimodal information (Ngiam et al., 2011). In practice, the analysis of disaster-related social media data is conducted by humanitarian organizations and crisis response agencies during both the early and ongoing phases of disaster events to obtain a summarized and operational view of evolving situations. Organizations such as the International Federation of Red Cross and Red Crescent Societies (IFRC), the Federal Emergency Management Agency (FEMA), and the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) actively employ social media analytics to support emergency response activities. For instance, IFRC and national Red Cross societies use social media streams to identify urgent needs, damaged infrastructure, and emerging crisis hotspots in near real time (Yu et al., 2020). FEMA integrates Twitter and other social media sources into situational awareness dashboards to facilitate rapid damage assessment, public communication, and inter-agency coordination during hurricanes and floods (Dereli et al., 2021; Madichetty & Muthukumarasamy, 2020). Similarly, UN OCHA applies social media–based information extraction and humanitarian data analytics to assist in needs assessment, international aid coordination, and the allocation of relief resources during large-scale disasters (Karami et al., 2019). Without automated analysis tools, critical information can be overlooked due to information overload, potentially delaying response efforts and increasing risks to affected populations.

Traditional approaches to disaster tweet classification often rely on unimodal data, processing text or images independently. Text-based models, such as those using BERT or Recurrent Neural Networks (RNNs), excel at capturing semantic context but lack visual evidence (Mansur et al., 2024; Vieweg et al., 2014). Similarly, image-based models, leveraging architectures like VGG-16 or ResNet, identify visual patterns but miss narrative context (Asif et al., 2021; Imran et al., 2020; Madichetty & Muthukumarasamy, 2020). Unimodal approaches remain widely adopted due to their lower computational cost, ease of deployment, and the availability of large labeled textual datasets. During disaster events, most social media posts contain only text or only images, which are easier to train, computationally efficient, and more suitable for real-time emergency response. However, by processing text or images in isolation, these methods fail to capture the complementary relationship between visual evidence and semantic context. As a result, they provide only a partial understanding of disaster events, which limits their effectiveness in accurately identifying informative content and supporting reliable situational awareness. Multimodal methods integrate multiple data sources, such as textual and visual information, enabling a more comprehensive analysis of disaster-related tweets. Text provides semantic cues such as disaster descriptions, locations, and urgency, while images offer visual evidence of damage severity, affected infrastructure, and on-ground conditions. Fusing modelling of these complementary modalities improves the identification of informative content and enhances situational awareness during disaster response.

Conventional fusion strategies, such as early and late fusion, rely on static weighting schemes and struggle to adapt to the highly dynamic and noisy nature of disaster-related social media data (Laya et al., 2021; Munawar et al., 2019). To address this limitation, recent studies have explored attention mechanisms (Sathianarayanan et al., 2024), which dynamically weight modalities based on contextual relevance. While attention-based methods have been widely adopted in general vision-language tasks, such as image captioning (Mozannar et al., 2018) and visual question answering (Bani Sakher et al., 2020), their application in disaster tweet classification remains limited. Motivated by this, proposes a cascaded attention-based multimodal framework specifically designed to classify disaster-related tweets as informative or non-informative. The framework employs self-attention to refine intra-modal features from text and images, and cross-attention to dynamically align inter-modal information. By integrating Bi-LSTM for text processing, VGG-16 for image analysis, and SMOTE to address class imbalance, the proposed approach outperforms unimodal and existing multimodal models. This framework supports real-time disaster response by enabling humanitarian organizations who often operate across regions and rely heavily on social media for rapid needs assessment and coordination to efficiently identify critical information and make timely, informed decisions.

The paper is organized as follows: Section 2 reviews related work, Section 3 discusses the proposed method, Section 4 details unimodal classification methods, Section 5 presents the multimodal classification of tweets, Section 6 describes the dataset and evaluation metrics, Section 7 details experimental results, Section 8 presents practical implications and deployment considerations, and Section 9 concludes with future directions.

2. RELATED WORK

Research on classifying disasters through social media has progressed over nearly a decade, leading to the development of various methods that employ different modalities. These methods can be classified into two primary categories based on modality types: (i) Unimodal Methods and (ii) Multimodal Methods.

2.1 Unimodal Methods

Unimodal methods rely on a single type of data or modality to identify disasters, such as text-based analysis, image recognition, or video content analysis. While the advantages of unimodal methods are their simplicity and focused approach, they may lack the comprehensive perspective provided by multimodal approaches. Textual tweets are the most basic form of data on which rigorous research has been done. Researchers have employed several machine-learning methods to explore textual data.

One study investigates flood-related tweets, categorizing them as high and low priority to predict high-priority tweet locations using a Markov model (Vaswani et al., 2017). A weakly supervised bag-of-words model refined via model transfer and updated online for real-time disaster tweet detection. Its effectiveness is limited by domain adaptability (Singh et al., 2017). The research highlights the value of combining shallow and deep learning for effective disaster tweet classification. Another study examines combining CNNs for feature extraction with ANNs for classifying tweets, enhancing disaster response by efficiently filtering vital information from large data streams (Palshikar et al., 2020). Self-learning algorithms build a simple, human-readable bag-of-words for processing disaster information with minimal supervision, using transfer learning to adapt the model to tweet-specific language (Sreenivasulu et al., 2019). Additional work has explored the challenges of detecting geolocation-based content communities on Twitter to mitigate the impact of disasters (Maswadi et al., 2024). Combining the strengths of CNN and GRU and incorporating the SkipCNN model improves the identification of crisis-related information derived from social media data (Nguyen et al., 2017). In another work, researchers underscore the potential of social media crowdsourcing in facilitating damage assessment in the aftermath of earthquakes by leveraging NLP and machine learning classifiers (Paul et al., 2023).

Equipping UAVs with cameras and transmitting real-time imagery to a base station enables a feature-concatenation deep learning model to classify disasters with higher accuracy than existing methods, enhancing disaster management in remote and inaccessible regions (*Li et al., 2023*). A study highlights the use of CNN for text classification and back-propagation neural networks, demonstrating how social media data enhances disaster management and economic loss assessments after typhoons (*Bashir et al., 2024*). In practice, several humanitarian organizations currently employ unimodal approaches**.** For instance, the American Red Cross monitors social media streams to detect posts reporting injuries, shelter needs, and urgent requests track keywords to generate flood maps for local responders. Platforms such as AIDR classify Twitter messages, enabling rapid prioritization by humanitarian partners. These

operational systems primarily rely on text-based models because they can be deployed in near real time, require less computational power, and integrate with existing monitoring dashboards. Although they are simple, unimodal methods lack the comprehensive perspective offered by multimodal approaches, limiting situational awareness.

2.2 Multimodal Methods

Multimodal methods enhance disaster classification by integrating inputs like text, images, and video for a more comprehensive understanding. While they improve accuracy and robustness, they are typically more complex and resource-intensive. The Crisis-DIAS framework involves a twofold task by combining attention mechanisms and a gated framework to merge text and image cues (Agarwal et al., 2020*).* A decision diffusion method combines text and image features using traditional deep learning models for crisis tweet classification, which outperformed unimodal approaches but lacks fine-grained cross-modal interaction (Gautam et al., 2019*).* Causal inference enhances disaster classification on social media by integrating textual and visual data, but its performance is constrained by noisy, unstructured inputs and high computational overhead, limiting its real-time applicability (Moraffah et al., 2022).

A graph-based multimodal framework that integrates textual, visual, and social context features using graph neural networks and attention mechanisms to improve disaster content classification (Dar et al., 2025). The DMCC framework integrates text and image data using feature and score-level fusion for improved crisis classification. However, its performance is limited by data quality, generalizability across crisis types, and high computational costs (Rezk et al.,2023). Contrastive learning models, CLIP and ALIGN assess the effectiveness of pre-trained multimodal models in classifying crisis-related tweets. However, these models assume a straightforward correlation between text and images, which may not hold for crisis-related tweets (Mandal et al., 2024). The MCAN model for Visual Question Answering employs modular co-attention layers that combine self-attention and guided attention to align textual and visual elements. Each layer builds on the previous, enhancing reasoning over both modalities (Yu et al., 2019). The DSACA model enhances VQA by combining dual self-attention for intra-modal dependencies and co-attention for cross-modal alignment. This creates a richer joint embedding of visual and textual features for improved reasoning (Liu et al., 2021). A causality-guided adversarial multimodal framework for crisis classification is introduced that improves generalization across unseen disasters by disentangling causal and spurious features and aligning text–image representations in a shared latent space (Ma et al., 2025). A cross-attention-based multimodal model significantly improves the classification of ambiguous disaster tweets, highlighting the value of attention-driven multimodal integration for humanitarian response (Teng & Öhman, 2025). From the above discussion, it is clear that there is a growing interest among researchers in utilizing deep multimodal frameworks for disaster-related posts on social media. While researchers have focused on early fusion and late fusion methods to integrate modalities. The specific challenges of this study are as follows:

- Traditional models using only text or visuals lack situational depth; we address this with cross-attention for dynamic text-image interaction. To address this, we integrate both modalities using a cross-attention mechanism that enables dynamic interaction between text and image features.
- Early and late fusion use fixed weights and fail to capture fine-grained inter-modal relationships. Our adaptive attention captures fine-grained inter-modal relevance for better performance.
- Direct concatenation can add noise. Instead of direct concatenation, self-attention filters relevant features, enhancing selection and reducing interference.
- The potential of attention mechanisms in multimodal disaster tweet classification is underexplored, motivating our investigation.

While attention mechanisms have proven effective in other domains, their potential in multimodal disaster classification remains underexplored, particularly in integrating textual and visual tweets for improved feature selection. This gap in research motivates us to investigate the potential of self and cross-attention mechanisms in enhancing multimodal approaches for the improved classification of disaster tweets. By addressing these challenges, the proposed attention-based multimodal classification model aims to improve the accuracy and efficiency of disaster tweet classification.

3. PROPOSED WORK

The proposed multimodal architecture for binary classification of disaster-related tweets integrates information from both textual and visual modalities corresponding to the same disaster event. As illustrated in Figure 1, the framework processes tweets containing both text and images posted during disaster situations. Visual features are extracted using a pre-trained VGG-16 model, while textual features are obtained using a two-layer Bi-LSTM that captures contextual representations for each word. To address the class imbalance in the dataset, particularly the underrepresentation of informative disaster tweets, SMOTE (Synthetic Minority Oversampling Technique) is applied. Imbalanced data can bias deep multimodal models toward the majority class. SMOTE generates synthetic samples for the minority class, enhancing its representation without merely duplicating existing data.

To effectively fuse the modalities, the proposed framework incorporates self-attention and cross-attention mechanisms. Self-attention is applied within each modality to capture intra-modal dependencies, and cross-attention enables the model to learn inter-modal interactions. The fused representation is used to classify tweets into *informative or non-informative* categories, effectively utilizing complementary information from both modalities. In the following subsections, we delve into the unimodal classification of textual and visual features into informative and non-informative categories of disaster tweets. These unimodal pipelines serve as foundational components for our multimodal framework and highlight the strengths and limitations of each modality when used in isolation.

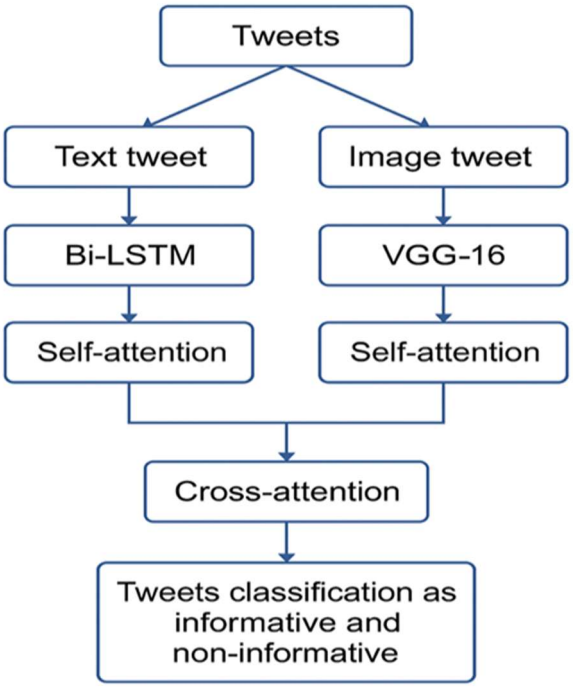

Figure 1. Overview of the proposed multimodal disaster tweet classification framework

4. UNIMODAL CLASSIFICATION OF DISASTER TWEETS

4.1 Unimodal Classification of Text

For textual features, we explore deep learning architectures such as Bi-LSTM and Bi-GRU (Schuster & Paliwal, 1997). This captures semantic and contextual nuances in tweets, making them useful for **classification** into *informative* and *non-informative* categories as shown in Figure 2. The goal is to determine whether the tweet is *informative,* containing actionable or critical information about a disaster, or *non-informative*, such as general commentary or unrelated content. These unimodal text classifiers provide a foundational understanding of textual relevance in disaster scenarios.

4.1.1 Preprocessing of Text

Text preprocessing prepares raw tweets for analysis, transforming them into a cleaner and more structured format. This process involves removing unwanted components, including URLs, mentions, special characters, punctuation, hashtags, emojis, and extra spaces. By cleaning tweets into structured textual input, preprocessing facilitates effective feature extraction, resulting in more accurate and meaningful analyses.

Following preprocessing, each word in the tweet is transformed into a dense, real-valued vector for further analysis. We employ pre-trained FastText embeddings, which extend traditional word embeddings by representing each word as a combination of character n-grams that capture both semantic meaning and morphological structure. This effectively handles out-of-vocabulary (OOV) words, a common challenge in noisy social media text. Such capability

is particularly advantageous for tweets containing misspellings, slang, or domain-specific terms. In our implementation, we utilize 200-dimensional FastText vectors to preserve rich linguistic information for downstream processing.

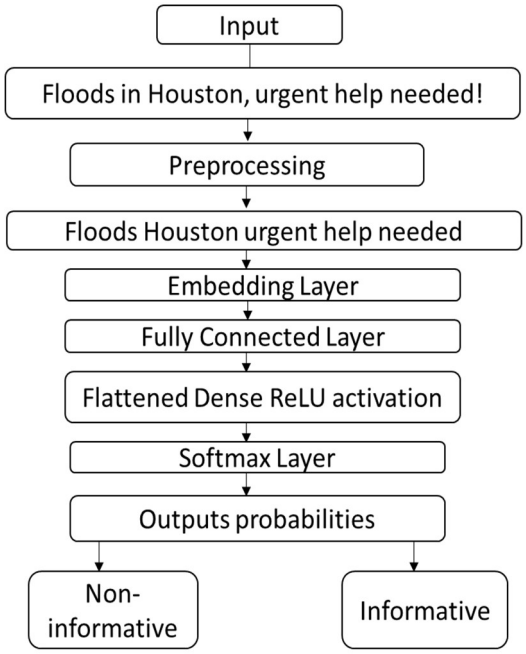

Figure 2. Unimodal text classification pipeline using Bi-LSTM/Bi-GRU for disaster-related tweets

4.1.2 Classification Using Bi-LSTM

Once the textual tweet is preprocessed, the Bi-LSTM network is employed to extract semantic features from the sequence of tweet tokens. The model utilises two LSTMs, one processes the input sequence in the forward direction, while the other processes it in reverse. This bidirectional structure allows the model to capture both past and future contextual dependencies within the tweet. At each time step t, the forward LSTM generates a hidden state $ht\rightarrow$, and the backwards LSTM generates $ht\leftarrow$. Let the input text tweet sequence be denoted as E = {e1, e2, …, en}, where n is the number of tokens. The text is embedded into 1024-dimensional vectors, and the Bi-LSTM encodes each token into a context-aware representation. The output feature matrix for a tweet is of shape (n,1024), with each row representing the hidden state of a token:

$$E= \{h_e \mid h_e \in R^d\} \quad (1)$$

Where d is the dimension of the hidden state, e is the length of the sentence m, and h_e corresponds to the hidden state of token e. These are concatenated to form the final hidden state:

$$h_t= [\rightarrow h_t, \leftarrow h_t] \quad (2)$$

4.1.3 Classification Using Bi-GRU

The Bidirectional Gated Recurrent Unit (Bi-GRU) extends the standard GRU by processing sequences in both forward and backward directions [39], enabling a richer contextual understanding of text tweets. Unlike standard GRUs, which operate unidirectionally, Bi-GRU employs two GRUs: the forward GRU processes the tweet from start to end, producing $ht\rightarrow$, while the backward GRU processes it in reverse, yielding $ht\leftarrow$. These are concatenated at each time step as per Equation (1).

Given an input tweet sequence $E = \{e_1, e_2, \ldots, e_n\}$, where n is the number of tokens, each token is embedded and passed through the Bi-GRU to generate a context-aware feature matrix. This bidirectional processing makes Bi-GRU particularly effective for understanding the full semantic structure of textual tweets.

4.2 Unimodal Classification of Images

Figure 3 shows the unimodal classification of disaster-related tweets that contain images, without incorporating textual data. Specifically, it explores the use of **VGG-16** and **ResNet-50** as the backbone architectures for image classification (He et al., 2016; Simonyan & Zisserman, 2014). VGG-16 offers a straightforward yet effective architecture, utilizing stacked convolutional layers, while ResNet-50 introduces residual connections that facilitate the capture of complex patterns and deeper representations. These features are then fed into fully connected layers to classify the image as informative or non-informative. This approach helps assess the standalone discriminative power of pictures in disaster tweet classification.

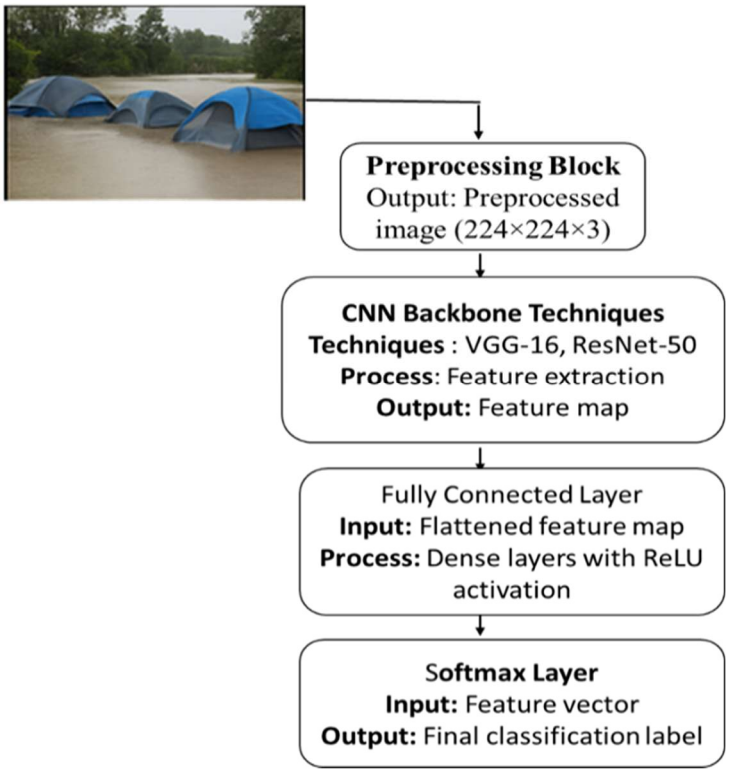

Figure 3. Unimodal image classification pipeline using CNN-based techniques

4.2.1 Preprocessing of Images

We begin by resizing the input image to a fixed dimension of 256x256 pixels to ensure consistency across all images, after which a random patch of 224x224 pixels is cropped from the resized image. Scaling pixel values to the range of [0,1] contributes to stabilizing and accelerating the training process, making the optimization more effective. Then, the pixel values are normalized using the mean and standard deviation from the ImageNet dataset.

4.2.2 VGG-16

VGG16 is widely recognized for its simplicity and effectiveness in image classification tasks. In this study, we adapt VGG16 to extract visual features from disaster-related image tweets for binary classification into informative and non-informative categories. The input image tweets are resized to the standard input dimension of:

$$I \in \mathbb{R}^{224\times224\times3} \tag{3}$$

The model comprises 13 convolutional layers, organized into five convolutional blocks, each employing 3×3 filters. These filters perform a convolutional operation on the image, performing element-wise multiplication followed by summation to generate feature maps, effectively capturing spatial patterns and textures across the image. To retain essential features and reduce the dimensionality of feature maps, VGG16 incorporates five max-pooling layers, each applying 2×2 pooling with a stride of 2. This compresses the spatial representation while preserving salient activations. After the convolutional and pooling stages, the architecture transitions to three fully connected dense layers. The first two fully connected layers each have 4096 units. The final layer comprises 1000 units used to perform binary classification. The dimensions of the output feature map are (7, 7, 512), where 7x7 denotes the spatial dimensions, and 512 denotes the number of feature channels. The feature map is reshaped as:

$$G\epsilon\mathbb{R}^{49\times512} \tag{4}$$

The 7×7 spatial grid is reshaped into 49 spatial regions, each represented by a 512-dimensional feature vector. This reshaped feature map is denoted as:

$$G=\{v_i|v_i\in R^{dv},i=1,2,3,\ldots.,M\} \tag{5}$$

Here, M denotes the total number of regions, which is 49, and vi is a 512-dimensional vector representing the features of regions. This step ensures image features are effectively extracted and converted into an appropriate format for subsequent steps in the model. Additionally, using the ReLU activation function throughout the convolutional layers accelerates convergence and enhances the model's ability to learn complex, non-linear patterns by preserving positive activations and eliminating negative ones.

4.2.3 Residual Network (ResNet-50)

The ResNet model revolutionized deep learning by addressing the challenges associated with training very deep networks, which increases the error rate. We utilized the ResNet-50 variant, which is extremely deep, where the number indicates the total number of layers. The use of

residual connections, or skip connections, allows the input image of a block to be added directly to the output produced by the last convolutional layer within that block. The architecture allows for the training of networks with thousands of layers. Each typical residual block is succeeded by batch normalization and ReLU activation functions. Each residual block includes three layers: a 1x1 convolution to reduce dimensionality, a 3x3 convolution kernel for feature extraction, and another 1x1 convolution to restore dimensionality. Residual blocks perform identity mapping; if no learning occurs, the resulting visuals remain identical to the original input.

5. MULTIMODAL CLASSIFICATION OF TWEETS

We propose a comprehensive multimodal framework designed to classify disaster-related tweets by effectively integrating both textual and visual information. As shown in Figure 4, each input tweet is represented as a multimodal pair C = (E, G), where E is the text tweet, and G is the corresponding image. To obtain a meaningful representation of the textual tweet, we extend the conventional Bi-LSTM model by incorporating self-attention, which converts the sequence of words into a feature matrix $T \in \mathbb{R}^{L\times d}$, where L is the number of tokens and d is the embedding dimension. Simultaneously, the image is processed using a self-attention-augmented VGG16, which allows the model to integrate global contextual relationships across the image to generate a visual feature matrix $I \in R^{N\times d}$, where N denotes the number of spatial regions, e.g., 49 from a 7×7 feature map and each region is mapped to the same embedding dimension d to enable fusion with the text features.

Applying self-attention separately to the text and image features captures intra-modal dependencies within each modality. Bi-LSTM effectively captures contextual dependencies in both forward and backward directions; its final representation often treats all tokens equally, potentially diluting the importance of disaster-critical keywords. To enhance the textual representation of disaster-related tweets, we extend the conventional Bi-LSTM encoder by incorporating a self-attention mechanism over its hidden states. This dynamically focuses on different parts of the same modality and captures local dependencies within each modality. For the textual embedding matrix E, the self-attention mechanism generates query (Q), key (K), and value (V).

$$Q_E = EW_Q \ , \ K_E = EW_K, \ V_E = EW_V \tag{6}$$

Here, W_Q, W_K, W_V they are learnable weight matrices. These matrices are then used to compute the attention output A_E using the scaled dot-product attention formula:

$$A_E = \left(softmax\left(Q_E \cdot K_E^T\right) \Big/ \sqrt{d_k}\right).V_E \tag{7}$$

Here, d_k is the dimension of the key vectors. This enables each word to attend to other contextually relevant words within the tweet. By integrating self-attention, the model can selectively focus on the most informative words, thereby producing a more discriminative textual feature vector. These enriched features improve cross-modal alignment with tweet text, resulting in superior classification of disaster-related social media content.

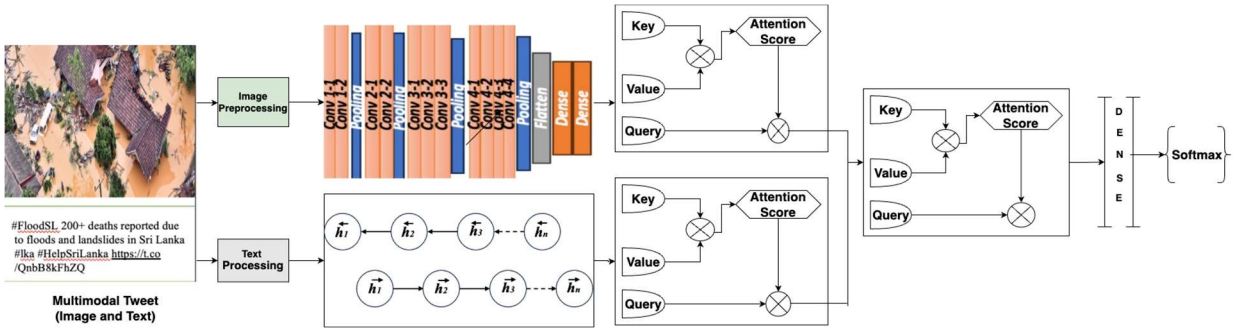

Figure 4. Proposed multimodal tweet classification framework

Similarly, to enhance contextual understanding for the visual features G, self-attention is applied to model relationships among different spatial regions of the image, dynamically emphasize disaster-relevant regions, and generate context-aware image representations.

$$Q_G = \mathrm{G}W_Q,\ K_G = \mathrm{G}W_K,\ V_G = \mathrm{G}W_V \tag{8}$$

$$A_G = (softmax(Q_G \cdot K_G^T) \Big/ \sqrt{d_k}).\ V_G \tag{9}$$

This enables the model to highlight salient regions by understanding intra-image dependencies on feature similarity in the context of disaster scenes, such as smoke, rubble, or fire.

To facilitate effective interaction between the modalities, we employ a cross-attention mechanism that allows features from one modality to attend to those of the other. Specifically, we use the image features as queries, and the text features as keys and values. This enables each image region to selectively attend to the most relevant parts of the text. Specifically, we compute textual and visual features through $\mathrm{E_T}$ and $\mathrm{G_I}$,

$$Q_E = \mathrm{C}W_Q, K_G = \mathrm{C}W_K, V_E = \mathrm{C}W_V \tag{10}$$

Using scaled dot-product attention, cross-modal interactions are modeled as:

$$A_{cross\_att} = (Q_{A_E} . K_{A_G}^T) \Big/ \sqrt{d_k} \tag{11}$$

$$A_{E,G} = A_{cross\_att} \odot V_{A_E} \tag{12}$$

This is the image-aware text representation, which means textual tokens are now enriched with visual information, where $\odot$ denotes element-wise multiplication. The result $A_{E,G}$ is a set of spatially-aware visual embeddings that attend to the most relevant parts of the text tweet,

enriched with contextual textual information. This enables each region of the image to “ask” what parts of the text describe it best. This fusion captures critical inter-modal interactions, such as associating the visual cue of smoke with the textual word “fire.” The fused representation $A_{E,G}$ is then flattened and passed through a series of fully connected layers equipped with the ReLU (Rectified Linear Unit) activation function. This is followed by a final softmax classification layer that outputs probabilities for each class.

$$A_{Dense}=\text{Dense}(A_{E,G}), A_{Final}=\text{Softmax}(A_{Dense}) \quad (13)$$

The output reflects the model’s prediction of whether the tweet is ‘informative’ or ‘non-informative’ with respect to disaster response relevance. The output $A_{Final}$of the classification layers provides predictions regarding whether the integrated visual-textual pair is 'informative' or 'non-informative'. This ensures that the model captures relevant interactions between the modalities for the classification task.

This end-to-end architecture ensures deep, meaningful interaction between textual and visual modalities. By combining self-attention for internal feature refinement and cross-attention for inter-modal alignment, the model can effectively relate textual cues such as "flooded," "collapsed," or "injured" with corresponding image regions. This significantly enhances the classification accuracy of multimodal disaster tweets, supporting better situational awareness and timely crisis response.

6. DATASET AND EVALUATION METRICS

In this section, we discuss the Crisis Multimodal Data (CrisisMMD) dataset, which is a benchmark dataset designed for crisis-related social media analysis that contains multimodal Twitter data collected during various natural disasters. The evaluation metrics compared the performance of the proposed framework against unimodal texts and images and two multimodal baseline models.

6.1 Dataset

The multimodal CrisisMMD Twitter dataset employed in the study (Ofli et al., 2020) fulfills the requirement of managing tweets comprising textual content and corresponding visual tweets from seven natural disasters. Each textual and visual tweet is annotated to assess its relevance for humanitarian efforts, thereby facilitating a clear differentiation between informative and non-informative content. The dataset includes tweets about the following natural disasters: Hurricane Harvey, Hurricane Irma, Hurricane Maria, the California Wildfires, the Mexico Earthquake, the Iraq-Iran Earthquake, and the floods in Sri Lanka.

This study classifies disasters into four specific categories: hurricanes, wildfires, earthquakes, and floods, as shown in Table 1. The dataset consists of 16,058 textual tweets and 18,082 visual tweets, which are categorized into two classes: informative and non-informative. Tweets which

deliver practical insights, including details about casualties, damage assessments, warnings, and requests for assistance, are informative. In contrast, non-informative tweets may convey sympathy or general sentiments and can include visual content of political figures or promotional content that lacks practical relevance for disaster management. The accompanying figure delineates the differences between informative and non-informative data in disaster response as shown in Figure 5, highlighting the significance of a multimodal framework for filtering and interpreting essential information. Since a single tweet can feature up to four images as shown in Figure 6, the frequency of image tweets often surpasses that of textual tweets. For analysis, only those tweets in which textual-visual pairs share identical labels are included. Since a single tweet can feature up to four images, the frequency of image tweets often surpasses that of textual tweets. For analysis, only those tweets in which textual-visual pairs share identical labels are included. The resulting filtered dataset comprises 12,762 tweets, with 8,463 labelled as informative.

Table 1. Dataset composition

Disaster	Total Tweets	Informative	Non-informative	Train set	Validation set	Test set
Hurricane	9075	6107	2968	5806	1453	1816
Wildfire	1205	923	282	771	193	241
Earthquake	1621	1204	417	1036	80	325
Flood	861	229	632	550	138	173

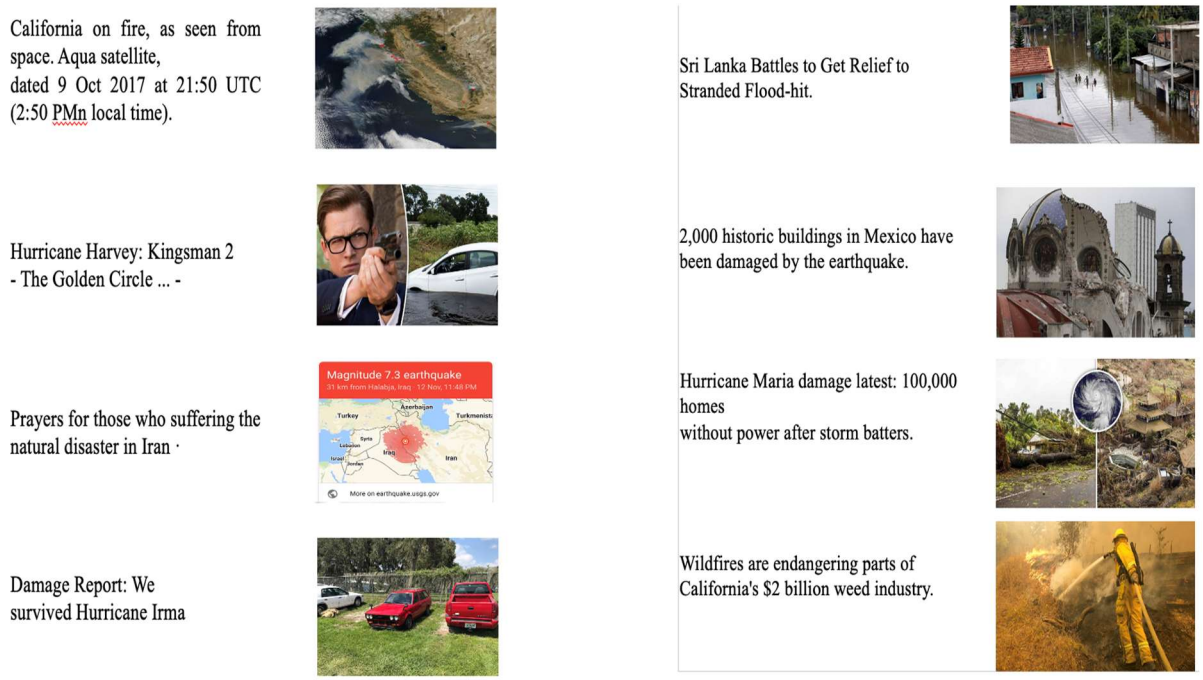

Figure 5. Non-informative visual and textual tweets (left), Informative visual and textual tweets (right)

6.2 Evaluation Metrics

In the present study, several evaluation metrics are utilized to assess the performance of trained models on the CrisisMMD dataset. These metrics provide insights into how well the

model classifies tweets into 'informative' and 'non-informative' categories based on their content:

- **Accuracy**: The most valuable evaluation metric is accuracy, which measures how accurately the model performed by taking the ratio of correctly classified tweets relative to the total number of tweets.

  Accuracy = True Positive tweets + True Negative tweets \ Total number of tweets
- **Precision**: Precision is the model's ability to correctly predict positive tweets relative to the total predicted positives.

  Precision = True Positive tweets/True Positive tweets + False Positive tweets
- **Recall**: Recall is the ratio of correctly predicted positive tweets (true positives) to the total actual positive samples.

  Recall=True Positive tweets /True Positive tweets + False Negative tweets
- **F1-score**: The F1-score is a single metric that combines the harmonic mean of precision and recall.

  F1-score=2×Precision×Recall/ Precision +Recall

Figure 6. Examples of textual tweets with multiple images

7. RESULTS AND DISCUSSION

To validate the efficacy of the proposed self and cross-attention multimodal framework, extensive experiments were carried out using the multimodal dataset, which includes data from seven different types of disasters. The performance of both unimodal and multimodal experiments is compared with the proposed method, which utilizes a novel attention mechanism. Traditional methods, including early fusion and late fusion, serve as baselines, alongside two multimodal models, MCAN and DSACA. The dataset was split into training, validation, and testing sets using an 80:10:10 ratio. Visual features were extracted using a VGG-16 network pre-trained on ImageNet, while textual sequences were modeled using a Bi-LSTM with dropout for regularization. The models were implemented in PyTorch and trained for 30 epochs using the Adam optimizer. Binary Cross-Entropy Loss was employed, with validation-based tuning to ensure robust performance.

7.1 Performance evaluation of unimodal models

This section elaborates on the performance of unimodal models. It explores text-based and image-based models. Bi-LSTM and Bi-GRU are text-based models, while VGG-16 and ResNet-50 are image-based models. The classification has been done according to four disaster types: hurricanes, wildfires, earthquakes, and floods. The models were assessed using accuracy, precision, recall, and F1-score to determine their effectiveness in disaster classification. For textual analysis, Bi-LSTM and Bi-GRU models show moderate performance, with weighted average F1-scores of 63.21% and 62.42%, respectively. Their performance varies across disaster types, reflecting the challenges of relying solely on short, noisy, and ambiguous text during crisis events. In particular, lower precision and recall for wildfire and earthquake tweets indicate difficulty in capturing context when textual cues are limited or unclear.

VGG-16 and ResNet-50 models are used for image-only disaster classification. Overall, VGG-16 achieves a higher weighted F1-score of 75.13% than ResNet-50 with 73.64%, indicating more stable performance across disaster types. VGG-16 performs particularly well for earthquake and flood events, where its simpler and more uniform convolutional structure appears effective in capturing coarse visual cues such as collapsed buildings and flooded regions. The results suggest that while deeper architectures like ResNet-50 can achieve high performance in specific scenarios, VGG-16 provides more consistent and robust visual representations for disaster-related imagery. This stability makes VGG-16 a suitable choice for integration within the proposed multimodal framework, where reliable visual features are essential for effective cross-attention fusion with textual representations.

Table 2 summarizes the performance of unimodal textual and visual models across different disaster types. For textual classification, the Bi-LSTM model consistently outperforms the Bi-GRU baseline, achieving an average absolute accuracy improvement of 2.97%, which indicates its stronger capability to capture long-range contextual dependencies in disaster-related tweets. Among visual models, VGG-16 yields superior performance compared to ResNet-50, with an

average absolute accuracy gain of 3.98%, demonstrating its effectiveness in extracting discriminative visual features from disaster imagery.

Table 2. Unimodal models performance metrics

Disasters	Unimodal															
	Textual								Visual							
	Bi-LSTM				Bi-GRU				VGG-16				ResNet-50			
	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Hurricane	77.8	74.23	69.55	71.8	78	73.31	63.44	68.1	88.32	69.8	74.56	72.3	84.6	70.54	69.41	73.2
Wildfire	72.4	55.6	54.8	55.2	65	53.3	53.44	53.4	77.67	75.34	70.11	62.1	80.66	74.55	77.89	72.6
Earthquake	76.9	60.81	70.65	65.5	69.4	61.45	63.42	62.4	83.44	77.8	67.8	72.1	75.44	70.12	71.58	70.8
Flood	68.08	67.13	72.15	69.6	70.8	69.8	67.8	68.8	79.54	74.56	75.38	69.2	72.33	77.43	68.58	72.7
Weighted average	73.8	64.44	66.79	65.5	70.8	64.47	62.03	63.4	82.24	74.38	71.96	69.2	78.26	73.16	71.87	72.1

Overall, visual unimodal models outperform textual ones across all evaluation metrics. In particular, VGG-16 achieves the highest weighted average accuracy of 82.24% and F1-score of 75.13%, highlighting the importance of visual cues for disaster understanding. However, despite their superior performance, visual-only approaches also face limitations, such as when images are missing, unclear, or unrelated to the disaster context. Overall, the results indicate that while unimodal approaches are effective and computationally simpler, they provide incomplete situational awareness when used in isolation. This performance gap motivates the need for multimodal fusion strategies that can jointly leverage complementary textual and visual information for disaster analysis.

7.2 Performance Evaluation of Multimodal Models

When transitioning from unimodal to multimodal models, Table 3 presents the performance of various multimodal approaches for disaster classification using Twitter, evaluating early fusion, late fusion, MCAN, DSACA, and the proposed method. The proposed method demonstrates superior performance, achieving the highest accuracy of 90.81% and an F1-score of 85.63%. This highlights the strength of its multimodal attention mechanism, which effectively enhances intra-modal features while dynamically aligning cross-modal information. To illustrate performance gains over existing models, relative improvement is used as the comparison metric. Early fusion delivers the weakest results, with relative improvements of 15.08% in accuracy and 30.75% in F1-score, indicative of its limitations due to noisy feature integration. Late fusion performs slightly better, offering 13.91% relative accuracy improvement and a 14.44% F1-score gain, suggesting that separately extracting features before

merging yields more reliable classifications. Baseline models like MCAN and DSACA show further enhancements, achieving relative accuracy gains of 8.89% and 9.23%, and F1-score improvements of 6.63% and 6.95%, respectively. While both leverage attention mechanisms, they fall short in refining intra-modal representations, a gap effectively addressed by the proposed framework. The proposed method achieves superior performance across all disaster types, recording an average F1-score of 85.63%, as illustrated in Figure 7(a). This demonstrates clear improvements from unimodal models to the multimodal approach, with the proposed framework delivering the highest F1-score overall.

Table 3. Performance comparison of multimodal methods

Disaster	Early Fusion				Late Fusion				MCAN				DSACA				Proposed Method			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Hurricane	79.2	80.76	68.8	64.85	81.43	78.45	75	77.96	84.33	80.76	78.9	81.24	85.67	81.49	81.6	79.45	91.66	78.9	91.55	82.33
Wildfire	80.02	76.15	72.1	67.88	76.89	79.85	81	75.45	81.26	78.65	72.1	75.67	80.23	79.76	79.6	78.65	90.23	85.76	89.56	87.59
Earthquake	77.52	72.41	67.7	66.09	80.76	76.21	74.1	75.9	85.98	83.49	79.4	81.87	84.69	83.87	82.9	80.49	88.69	84.87	87.49	85.94
Flood	78.9	76.54	64.3	63.12	79.8	71.94	72.7	69.96	81.99	79.65	82.3	82.44	82.66	79.64	80.7	81.65	92.66	87.64	86.53	86.65
Weighted	78.91	76.47	68.2	65.49	79.72	76.61	75.7	74.82	83.39	80.64	78.1	80.31	83.31	81.19	81.2	80.06	90.81	84.29	88.78	85.63

Additionally, Figure 7(b) highlights its leading performance in terms of average accuracy, reaching a weighted accuracy of 90.81%. When broken down by disaster type, the F1-scores are: Hurricane 82.33%, Earthquake 85.94%, Wildfires 87.59%, and Floods 86.65%, reinforcing the effectiveness of the multimodal fusion strategy. Overall, multimodal attention mechanisms significantly enhance disaster classification, with the proposed method consistently outperforming other techniques. The choice of fusion strategy significantly impacts effectiveness, as attention-based approaches have proven superior to simple early fusion. These findings underscore the importance of advanced fusion mechanisms assessment and suggest that future research should focus on refining attention models and incorporating additional data sources for further improvements. Experiments on the CrisisMMD dataset

evaluate the proposed cascaded attention-based multimodal framework against unimodal and multimodal baselines across four disaster types: hurricanes, wildfires, earthquakes, and floods. Performance metrics include accuracy, precision, recall, and F1-score, providing a comprehensive assessment of classification effectiveness.

7.3 Qualitative Analysis

To provide a clear and effective evaluation of the proposed method on the CrisisMMD dataset, it's essential to examine specific examples where the model's predictions are juxtaposed with the expected true labels. This comparison highlights both the approach's strengths and potential areas for improvement, offering insight into the performance on different tasks. To effectively evaluate the proposed framework on the CrisisMMD dataset, it is crucial to compare model predictions with actual labels. Figure 7 illustrates accurate classifications of informative and non-informative tweets, showcasing the model's strength in fusing textual and visual data.

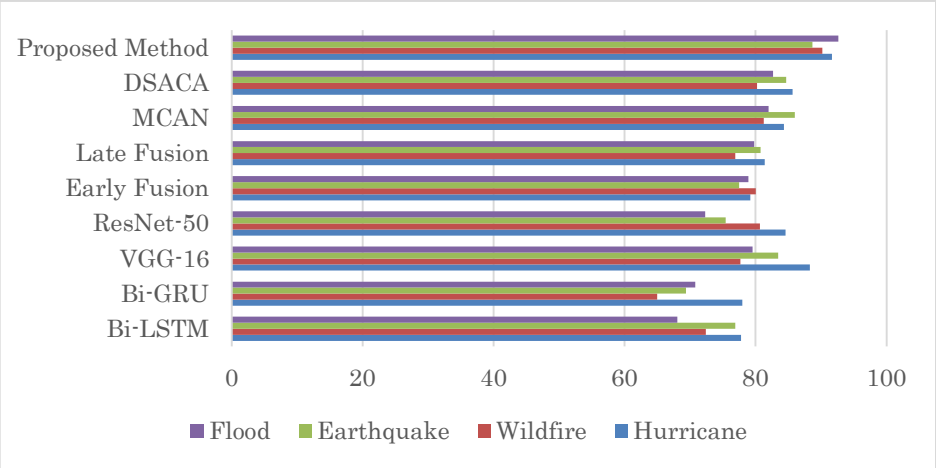

Figure 7(a). Model-wise average F1-score

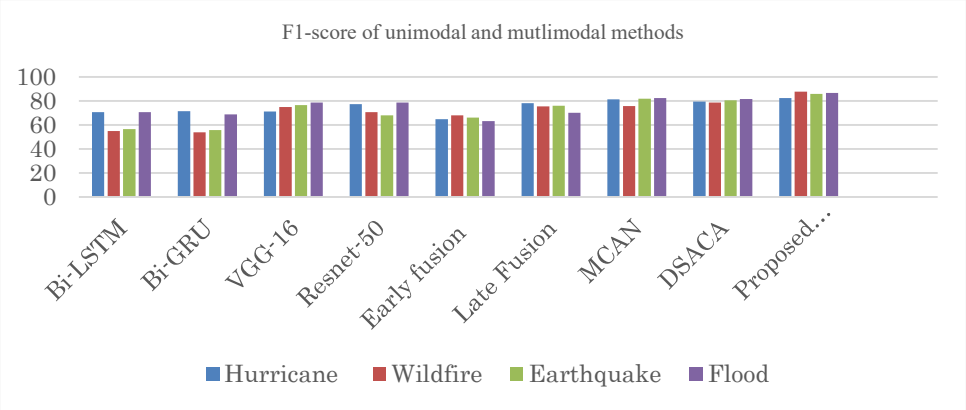

Figure 7(b). Compares the accuracy of all the disasters for the models

Figure 8(left) shows cases where the classifiers correctly identify tweets as informative and non-informative, showcasing the model's strength in fusing textual and visual data. These successful predictions demonstrate that the proposed multimodal fusion technique effectively

and proficiently integrates information from both textual and visual data to make accurate predictions. However, there are also instances of misclassification due to noisy visual features. Figure 8(right) shows tweets incorrectly predict a non-informative tweet as informative and an informative tweet as non-informative. These errors emphasize the need for improved preprocessing to reduce visual noise and enhance model reliability in real-world scenarios.

Figure 8. Predicted and actual labels are the same (left), Predicted and actual labels are not the same (right)

7.4 Statistical Significance Analysis

To assess whether the observed performance differences are statistically meaningful, paired t-tests were conducted. Table 4 demonstrates paired t-test results comparing unimodal baselines and the proposed multimodal method across four CrisisMMD disaster types using F1-scores. Bi-LSTM marginally outperformed Bi-GRU in textual classification, where p = 0.219.

Table 4. Statistical comparison of unimodal baselines and the proposed multimodal method using F1-scores on the CrisisMMD dataset

Comparison	Models Compared	t-statistic	p-value	Mean Difference (%)	Significance
Textual Unimodal	Bi-LSTM vs. Bi-GRU	1.55	0.219	0.79	Not significant
Visual Unimodal	VGG-16 vs. ResNet-50	0.47	0.668	1.48	Not significant
Cross-Modality	VGG-16 vs. Bi-LSTM	2.46	0.091	11.91	Marginal
Multimodal vs. Visual	CAMM vs. VGG-16	3.85	0.018	10.81	Significant
Multimodal vs. Textual	CAMM vs. Bi-LSTM	6.72	0.003	22.72	Highly Significant

Similarly, VGG-16 and ResNet-50 showed comparable performance for visual classification with $p = 0.668$. In contrast, visual models significantly outperformed textual models overall, with VGG-16 achieving higher F1-scores than Bi-LSTM with a p-value of 0.091. Most importantly, the proposed multimodal method achieved significant improvements over the best visual baseline. VGG-16 with a p-value of 0.018 and the best textual baseline, Bi-LSTM, $p = 0.003$, demonstrating the effectiveness of multimodal fusion for crisis tweet classification.

These results indicate consistent performance gains achieved through multimodal fusion across different disaster categories.

7.5 Ablation Study

This study conducts an ablation analysis of attention mechanisms in multimodal methods integrating textual and visual modalities. Self-attention captures intra-modality features but proves less effective for cross-modal interaction. Cross-attention captures interactions between the two modalities by considering one modality in the context of the other, enhances inter-modality integration and yields better performance, albeit with increased complexity. The findings from the hurricane dataset are summarized in Table 5.

Table 5. Effectiveness of attention mechanisms

SA	CA	Accuracy	Precision	Recall	F1-score
…		84.93	74.04	79.80	76.80
	…	86.45	77.76	80.80	80.98
…	…	90.65	79.80	87.65	84.56

The proposed model incorporates both self-attention and cross-attention, yielding the best overall performance and effectively capturing both intra- and inter-modality features. Although the integration of the attention mechanisms increases the total number of parameters, the model demonstrates substantial performance enhancement. Overall, the study underscores the critical role of cross-attention in multimodal learning. Although complex models require more parameters, they can deliver significantly better performance.

7.6 Limitations and Challenges

Despite its effectiveness, the proposed cascaded attention-based multimodal framework has several limitations. First, the model relies on the availability of paired text–image tweets, which restricts its applicability since many disaster-related tweets contain only text or only images. Another limitation is that the use of self-attention on both Bi-LSTM and VGG-16 features, followed by cross-attention fusion, increases computational complexity and may slow down large-scale or real-time analysis. Finally, errors introduced in the early attention stages can propagate to later fusion layers, particularly when dealing with noisy and unstructured social media data, making it more challenging for humanitarian analysts to clearly understand the model's decision process.

8. PRACTICAL IMPLICATIONS AND DEPLOYMENT CONSIDERATIONS

The operational value of social media in disaster response has been demonstrated in several real-world events. Facebook Disaster Maps, developed by Meta Platforms, uses aggregated and

privacy-preserving data to generate insights on evacuations, displacement, connectivity loss, and infrastructure disruption after disasters (Maas et al., 2019). This system converts large-scale platform activity into structured crisis intelligence and complements traditional damage assessment and rapid needs analysis. During the 2019 floods in Poland, platforms such as Facebook and Twitter supported multiple stages of disaster management (Domalewska, 2019). Authorities shared warnings and operational updates. Media organizations reported real-time developments. Citizens posted local conditions and personal experiences, and social media acted as an important supplementary communication channel. In Hurricane Harvey, Twitter played a critical role in emergency coordination. Many residents posted rescue requests and shared their locations when emergency lines were overloaded (Zou et al., 2023). Despite these, much prior work relies on aggregated statistics or manual qualitative review. The proposed method performs fine-grained tweet-level classification. It jointly analyzes text and images. Cross-attention mechanisms help verify textual claims with visual evidence. This reduces false positives and increases confidence in actionable content, thereby enabling more reliable real-time decision support for emergency responders.

9. CONCLUSION AND FUTURE WORK

The rapid dissemination of disaster-related information on social media, including text, images, and metadata, offers unprecedented opportunities for real-time crisis response. This paper presents a novel cascaded attention-based multimodal framework that integrates textual and visual modalities to classify tweets as informative or non-informative. By leveraging Bi-LSTM for text processing, VGG-16 for image analysis, and a combination of self- and cross-attention mechanisms, the model dynamically captures both intra- and inter-modal relationships, thereby overcoming the limitations of traditional fixed-weight fusion strategies. Evaluated on the CrisisMMD dataset, the proposed framework achieves an accuracy of 90.81%, precision of 84.29%, recall of 88.78%, and F1-score of 85.63%, surpassing unimodal baselines and state-of-the-art multimodal models MCAN and DSACA by significant margins. These results demonstrate its ability to accurately identify critical disaster-related information, enhancing situational awareness for humanitarian agencies. This framework has helped in improving disaster response and management. It will depict informative content related to reports of casualties, infrastructure damage, or urgent needs. Humanitarian agencies can prioritise their action based on the informative tweets. By enabling rapid identification of actionable information, the system can support timely resource allocation and targeted relief efforts, particularly in fast-evolving disaster scenarios. By doing so, it supports faster resource allocation and targeted relief in fast-evolving disaster scenarios. Future research will extend the framework to incorporate audio and video modalities, further enriching multimodal analysis. Additionally, exploring transfer learning and domain adaptation will address challenges in data-scarce disaster scenarios, ensuring robust performance for emerging crises. By advancing attention-based multimodal learning, this work paves the way for more effective disaster response and resource allocation, ultimately saving lives.

DATA AVAILABILITY

The data used to support the findings of this study are available from the corresponding author upon request.

FUNDING

This work was carried out without funding from any agency in the public, commercial, or not-for-profit sectors.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

During the preparation of this work, the author(s) used ChatGPT, developed by OpenAI, to enhance the readability and clarity of the text (e.g., improving grammar, sentence flow, and language polishing). After using this tool/service, the author(s) reviewed and edited the content generated by the tool as necessary and take(s) full responsibility for the content of the publication.

REFERENCES

Acerbo, F. S., & Rossi, C. (2017). Filtering informative tweets during emergencies: A machine learning approach. In *Proceedings of the 1st Context Workshop on ICT Tools for Emergency Networks and Disaster Relief*.

Agarwal, M., Leekha, M., Sawhney, R., & Shah, R. R. (2020). Crisis-DIAS: Multimodal damage analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 346–353. https://doi.org/10.1609/aaai.v34i01.5369

Asif, A., Khatoon, S., Hasan, M. M., Alshamari, M. A., Abdou, S., Elsayed, K. M., & Rashwan, M. (2021). Automatic analysis of social media images to identify disaster type and infer appropriate emergency response. *Journal of Big Data, 8*(1), 83. https://doi.org/10.1186/s40537-021-00464-4

Bani Sakher, M., Omar, M., Hong, S., & Adams, J. (2020). A human-centric approach to data fusion in post-disaster management. *Journal of Business Management Sciences, 8*(1), 12–20.

Bashir, M. H., Ahmad, M., Rizvi, D. R., et al. (2024). Efficient CNN-based disaster event classification using UAV images. *Neural Computing and Applications, 36*, 10599–10612. https://doi.org/10.1007/s00521-024-09610-4

Dar, S. S., Rehman, M. Z. U., Bais, K., Haseeb, M. A., & Kumar, N. (2025). Social context-aware graph-based multimodal learning for disaster classification. *Expert Systems with Applications, 259*, 125337. https://doi.org/10.1016/j.eswa.2024.125337

Dereli, T., Eligüzel, N., & Çetinkaya, C. (2021). Content analysis of International Federation of Red Cross and Red Crescent Societies tweets using machine learning techniques. *Natural Hazards, 106*, 2025–2045. https://doi.org/10.1007/s11069-021-04527-w.

Devaraj, A., Murthy, D., & Dontula, A. (2020). Machine-learning methods for identifying social media-based requests for urgent help during hurricanes. *International Journal of Disaster Risk Reduction, 51*, 101757. https://doi.org/10.1016/j.ijdrr.2020.101757

Domalewska, D. (2019). The role of social media in emergency management during the 2019 flood in Poland. *Security and Defence Quarterly, 27*(5), 32–43.

Gautam, A. K., Misra, L., Kumar, A., Misra, K., Aggarwal, S., & Shah, R. R. (2019). Multimodal analysis of disaster tweets. In *Proceedings of IEEE BigMM* (pp. 94–103). https://doi.org/10.1109/BigMM.2019.00-38

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR*. https://doi.org/10.1109/CVPR.2016.90

Imran, M., Alam, F., Qazi, U., Peterson, S., & Ofli, F. (2020). Rapid damage assessment using social media images by combining human and machine intelligence. *arXiv*. https://arxiv.org/abs/2004.06675

Karami, A., Shah, V., Vaezi, R., & Bansal, A. (2019). Twitter speaks: A case of national disaster situational awareness. *Journal of Information Science*. https://doi.org/10.1177/0165551519828620

Laya, M., Delimayanti, M. K., Mardiyono, A., Setianingrum, F., Mahmudah, A., & Anggraini, D. (2021). Classification of natural disasters on online news data using machine learning. In *Proceedings of the 5th International Conference on Electrical, Telecommunication and Computer Engineering* (pp. 42–46). https://doi.org/10.1109/ELTICOM53303.2021.9590125

Li, S., Wang, Y., Huang, H., et al. (2023). Typhoon disaster assessment using social media and neural networks. *Natural Hazards, 116*. https://doi.org/10.1007/s11069-022-05754-5

Liu, Y., Zhang, X., Zhang, Q., Li, C., Huang, F., Tang, X., & Li, Z. (2021). Dual self-attention with co-attention networks for visual question answering. *Pattern Recognition, 117*, 107956.

Ma, P., Zhao, C., Jiang, B., Vishnubhatla, S., Jeong, U., Beigi, A., Raglin, A., & Liu, H. (2025). CAMO: Causality-guided adversarial multimodal domain generalization for crisis classification. In Proceedings of the AAAI Conference on Artificial Intelligence, Volume(Issue), pages.

Maas, P., Iyer, S., Gros, A., Park, W., McGorman, L., Nayak, C., & Dow, P. A. (2019). Facebook disaster maps: Aggregate insights for crisis response and recovery. In *Proceedings of the ACM SIGKDD Conference*.

Madichetty, S., & Muthukumarasamy, S. (2020a). Detection of situational information from Twitter during disasters using deep learning models. *Sādhanā, 45*(1), 270.

Madichetty, S., & Muthukumarasamy, S. (2020b). Classifying informative and non-informative tweets by adapting image features during disasters. *Multimedia Tools and Applications, 79*(39–40), 28901–28923. https://doi.org/10.1007/s11042-020-09343-1

Mandal, B., Khanal, S., & Caragea, D. (2024). Contrastive learning for multimodal crisis tweet classification. In *Proceedings of the ACM Web Conference 2024* (pp. 4555–4564). https://doi.org/10.1145/3589334.3648143

Mansur, A. V., Langhorn, G., & Nelson, D. R. (2024). Exploring social contracts of disaster risk through Twitter narratives during a major storm. *Nature-Based Solutions, 6*, 100197. https://doi.org/10.1016/j.nbsj.2024.100197

Maswadi, K., Alhazmi, A., Alshanketi, F., et al. (2024). Empirical study of tweet classification systems for disaster response. *Journal of Ambient Intelligence and Humanized Computing, 15*, 3303–3316. https://doi.org/10.1007/s12652-024-04807-w

Moraffah, R., Unni, S. J., Raglin, A., & Liu, H. (2022). Causal data fusion for multimodal disaster classification. In *SBP-BRiMS Proceedings*.

Mozannar, H., Rizk, Y., & Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. In *Proceedings of the 15th ISCRAM Conference*.

Munawar, H. S., Hammad, A., Ullah, F., & Ali, T. H. (2019). After the flood: Image processing and machine learning for post-flood disaster management. In *Proceedings of the 2nd International Conference on Sustainable Development in Civil Engineering* (pp. 5–7).

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning* (pp. 689–696).

Nguyen, D., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data using CNNs. In *Proceedings of ICWSM* (Vol. 11, pp. 632–635).

Ofli, F., Alam, F., & Imran, M. (2020). Analysis of social media data using multimodal deep learning for disaster response. In *ISCRAM Proceedings*.

Palshikar, G. K., Apte, M., & Pandita, D. (2020). Weakly supervised and online learning of word models for disaster tweet classification. *Information Systems Frontiers, 20*(5), 949–959.

Paul, N. R., Sahoo, D., & Balabantaray, R. C. (2023). Classification of crisis-related Twitter data using deep learning. *Multimedia Tools and Applications, 82*. https://doi.org/10.1007/s11042-022-12183-w

Rezk, M., Elmadany, N., Hamad, R. K., & Badran, E. F. (2023). Categorizing crises from social media feeds via multimodal channel attention. *IEEE Access, 11*, 72037–72049. https://doi.org/10.1109/ACCESS.2023.3294474

Sathianarayanan, M., Hsu, P.-H., & Chang, C.-C. (2024). Extracting disaster location identification from social media images using deep learning. *International Journal of Disaster Risk Reduction, 104*, 104352. https://doi.org/10.1016/j.ijdrr.2024.104352

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing, 45*(11), 2673–2681. https://doi.org/10.1109/78.650093

Seneviratne, K., Nadeeshani, M., Senaratne, S., & Perera, S. (2024). Use of social media in disaster management: Challenges and strategies. *Sustainability, 16*(11), 4824. https://doi.org/10.3390/su16114824

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*.

Singh, J. P., Dwivedi, Y. K., Rana, N. P., Kumar, A., & Kapoor, K. (2017). Event classification and location prediction from tweets during disasters. *Annals of Operations Research, 283*, 737–757.

Sreenivasulu, M., & Sridevi, M. (2019). Detecting informative tweets during disasters using deep neural networks. In *Proceedings of the International Conference on Communication Systems & Networks*.

Teng, S., & Öhman, E. (2025). Using multimodal models for informative classification of ambiguous tweets in crisis response. In Proceedings of the 5th International Conference

on Natural Language Processing for Digital Humanities (pp. 265–271). Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5999–6009).

Vieweg, S., Castillo, C., & Imran, M. (2014). Integrating social media communications into the rapid assessment of sudden onset disasters. In *Lecture Notes in Computer Science* (Vol. 8851). https://doi.org/10.1007/978-3-319-13734-6_32

Yu, M., Huang, Q., Qin, H., Scheele, C., & Yang, C. (2020). Deep learning for real-time social media text classification for situation awareness: Hurricanes Sandy, Harvey, and Irma. In *Social sensing and big data computing for disaster management* (pp. 33–50). Routledge.

Yu, Z., Yu, J., Cui, Y. H., Tao, D., & Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6281–6290).

Zou, L., Liao, D., Lam, N. S. N., Meyer, M. A., Gharaibeh, N. G., Cai, H., Zhou, B., & Li, D. (2023). Social media for emergency rescue: Rescue requests during Hurricane Harvey. *International Journal of Disaster Risk Reduction, 85*, 103513.