International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 5, October, 2023, pp. 1263-1272
Journal homepage: http://iieta.org/journals/ria

Effective Disaster Management Through Transformer-Based Multimodal Tweet Classification

Gundabathina JayaLakshmi[1], Abburi Madhuri[2], Deepak Vasudevan[3], Balamuralikrishna Thati[4], Uddagiri Sirisha[2], Surapaneni Phani Praveen[2*]

[1] Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada 520007, India
[2] Department of Computer Science and Engineering, P V P Siddhartha Institute of Technology, Vijayawada 520007, India
[3] Department of Computer Science and Engineering, Koneru Lakshmaiah Educational Foundation, Vaddeswaram 522302, India
[4] Department of Computer Science and Engineering, Dhanekula Institute of Engineering & Technology, Vijayawada 521139, India

Corresponding Author Email: phani.0713@gmail.com

https://doi.org/10.18280/ria.370519

Received: 17 June 2023
Revised: 6 September 2023
Accepted: 11 September 2023
Available online: 31 October 2023

Keywords:
disaster tweet classification, multimodal data fusion Bi-LSTM Attention, RoBERTa vision transformer

ABSTRACT

The role of social media in crisis response and recovery is becoming increasingly prominent due to the rapid progression of information and communication technologies. This study introduces a transformative approach to extract valuable information from the enormous volume of user-generated content on social media, specifically focusing on tweets that can significantly aid emergency response and recovery efforts. The identification of informative tweets allows emergency personnel to gain a more comprehensive understanding of crisis situations, thereby facilitating the deployment of more effective recovery strategies. Previous studies have largely focused on either the textual content or the accompanying visual elements within tweets. However, evidence suggests a complementary relationship between text and visuals, offering an opportunity for synergistic insights. In response to this, a novel deep learning framework is proposed, which concurrently analyses both textual and visual components extracted from user-generated tweets. The central architecture integrates established methodologies, including RoBERTa for text analysis, Vision Transformer for image understanding, Bi-LSTM for sequence processing, and an attention mechanism for context awareness. The innovation of this approach lies in its emphasis on multimodal fusion, introducing rank fusion techniques to effectively combine the strengths of textual and visual inputs. The proposed methodology is extensively tested across seven diverse datasets, representing various natural disasters such as wildfires, hurricanes, earthquakes, and floods. The experimental results demonstrate a superior performance of the proposed system, compared to several existing methods, with accuracy levels ranging from 94% to 98%. These findings underscore the efficacy of the proposed deep learning classifier in leveraging interactions across multiple modalities. In summary, this study contributes to disaster management by promoting a comprehensive approach that exploits the potential of multimodal data, thereby enhancing decision-making processes in emergency scenarios.

1. INTRODUCTION

Disasters have the potential to strike unexpectedly and at any location, resulting in significant damage and unpredictability. These events can stem from both human-driven factors such as industrial accidents, conflicts, and riots, as well as natural phenomena like earthquakes and droughts. The repercussions of such disasters extend to various aspects of life, impacting individuals, the environment, and the economy. Swift and effective responses are crucial in mitigating the consequences of unforeseen emergencies. Timely actions for restoration and recovery play a pivotal role in the field of disaster management. Any delays in these efforts can lead to worsened outcomes.

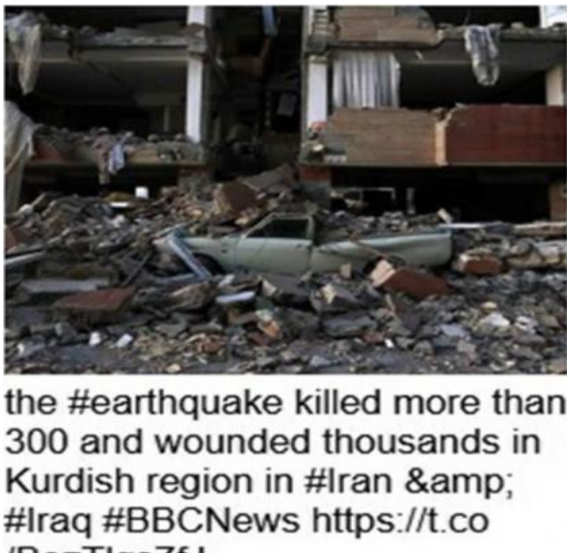
The realm of crisis informatics, as explored in the study [1], revolves around leveraging computational and informational sciences to facilitate disaster response. It serves as a vital bridge connecting individuals, devices, institutions, and data during critical situations. Recognizing the power, intelligence, and self-organizing capacity of the public, crisis informatics acknowledges the public's potential to influence emergency management outcomes significantly. A cornerstone of this field is the reliance on data, where effective decision-making hinges on access to timely and relevant information.

The rapid proliferation of social media platforms is predominantly attributed to technological advancements. These platforms offer real-time access to crucial information, and in today's news landscape, social media often kickstarts the dissemination of news and updates, as noted by Martínez-Rojas et al. [2].

In the realm of disaster informatics, information stands as the vital element. The surge of social media's popularity can be traced back to advancements in information and communication technologies, enabling the real-time provision of data. Notably, many significant news stories find their initial reporting on social media platforms [2].

Yet, the present era of prolific data production poses numerous challenges to the analysis of social media content. Within the expansive realm of social media, there exists a prevalence of conflicting information. Given the sheer volume of messages generated during a crisis, it becomes impractical for rescue personnel to keep pace with the influx of data. This deluge of messages further compounds the difficulty of identifying those that hold true urgency. The succinct and divergent nature of language used in social media posts adds complexity to the extraction of meaningful insights. Incorporating photographs and videos captured at disaster scenes can greatly enhance the comprehension of the situation. It has been established that relying solely on a single modality for investigation yields suboptimal outcomes, as previously underscored by researchers [3-8].

Historically, studies exploring social media have predominantly focused on textual or visual data in isolation [9]. However, a remarkable shift in understanding occurs when these two modalities converge. The fusion of image and text data presents a wealth of untapped insights, marking a promising frontier in the realm of multimodal data processing. Leveraging multiple modalities offers a more robust learning environment, underpinned by the contextual richness they provide. Thus, a pressing need emerges for an automated data processing system adept at discerning disaster-related keywords and concepts within tweets. A glimpse of recent crisis-related tweets (and accompanying images) is illustrated in Figure 1.



the #earthquake killed more than 300 and wounded thousands in Kurdish region in #Iran &amp; #Iraq #BBCNews https://t.co /BozTIqs7fJ

Figure 1. Sample multimodal tweets

The objective of this study is to establish a classification system capable of categorizing tweets based on both the textual content and any accompanying images. It's imperative for authorities to possess a comprehensive understanding of crucial information post-disaster to facilitate timely and effective recovery efforts.

To address multimodality, model integration is employed at both the output and feature levels [10]. This encompasses two approaches: Late Fusion, involving decision-level fusion, and Early Fusion, operating at the feature level, where features are combined and processed.

In our investigation, we delved into the dynamics of text and image features, exploring their interplay through the lens of text and image models. Leveraging advanced transformer-based models, we subjected both text and images to thorough processing. An innovative approach, termed early fusion, was introduced to discern interactions within input modalities.

Although the initial findings are promising, there is room for refinement in the proposed approach. Enhancing the system's capacity can be achieved through hardware upgrades, as deep learning is widely employed in various domains such as image captioning [11], object detection [12], and natural language processing, demanding substantial computational and memory resources, which may limit experimentation.

The method introduced here holds the potential for enhancing disaster response and recovery situational awareness, thereby enhancing overall quality of life. The implications of our findings extend to domains such as "fake news identification" and "question-answering systems", where diverse inputs are required.

Outlined below are the contributions offered by this paper:

-Texts and images undergo processing using transformer-based multimodal fusion systems.

-We propose a robust neural network architecture rooted in deep learning, featuring an innovative multi-modal feature fusion layer. It harnesses state-of-the-art deep learning techniques including BiLSTM, ViT model, and RoBERTa model. This approach capitalizes on the concealed interactions across diverse modalities, automatically detecting significant disaster-related tweets for responsive systems.

The structure of this paper unfolds as follows: In Section 2, we provide a succinct overview of pivotal works concerning catastrophic events. Section 3 delves into the comprehensive construction of a multimodal fusion system. The subsequent section, Section 4, outlines the system's approach and architecture. The experimental setup is expounded upon and evaluated in Section 5. Finally, in Section 6, we delve into an in-depth discussion of the model's outcomes.

2. RELATED WORK

Examination can be approached through various avenues, encompassing text-only, image-only, and multimodal analyses. Multimodal systems are versatile, capable of accommodating a diverse range of data sources, such as inputs from meteorological agencies and personal observations. The proposed methodology specifically capitalizes on multimodal inputs from Twitter, offering swift access to data during critical situations without the need for additional context. Numerous studies have demonstrated the valuable contribution of tweets containing both text and images in aiding disaster recovery efforts.

2.1 Text-based classification in disaster tweet analysis

Madichetty and Sridevi [13] employed a random forest classifier to discern tweets associated with damage. Their approach incorporated lexical, syntactic, and frequency properties of damage-related words. To enhance feature relevance, they employed "linear regression" and "support vector regression". Their model achieved a 94% earthquake prediction accuracy for Italy, Chile, and India. In the study

[14], a stacked CNN was recommended for identifying tweets concerning resource availability during crises. Crisis-specific terms were embedded, and the authors combined outputs from a "K-nearest neighbor" and a CNN classifier. The fused output underwent classification. In the same study, damage-indicative tweets were identified through a random forest classifier. The methodology leveraged lexical, syntactic, and frequency features, yielding a 94% earthquake prediction accuracy for the mentioned countries.

For tweet recognition under resource constraints and high accuracy, Madichetty [14] proposed a layered CNN. The study incorporated crisis-related vocabulary and merged outcomes of a CNN classifier with those of a K-nearest-neighbor classifier. Classification of the combined data utilized an SVM classifier, resulting in disaster dataset accuracies ranging from 67% to 77%. Snyder et al. [15] explored iterative connections with a system, enabling user intervention to rectify classifier errors. Emphasizing the tweet's content, the study employed convolutional neural networks, long short-term memory, and word2vec embedding for categorization based on severity.

Madichetty and Sridevi [16] proposed a majority voting-based ensemble classifier was developed to identify tweets containing links to pertinent medical resources with 82.4% accuracy, catering to those in need of medical assistance. Zahra et al. [17] employed a random classifier integrating linguistic features to distinguish eyewitness messages, employing sensory words, first-person pronouns, and unique adjectives. Utilizing the "distributional hypothesis" from linguistics, Ghafarian and Yazdi [18] proposed a novel approach was introduced, modeling tweets as word clouds and predicting label distribution via an SVM classifier. This concept outperformed the Bag-of-Words paradigm, demonstrating accuracies ranging from 74% to 80% across multiple datasets. Kejriwal and Zhou [19] proposed a minimally-supervised technique for identifying urgent disaster-related messages, leveraging both annotated and unannotated tweets to train the system and adapt to evolving emergencies.

2.2 Image-based classification in disaster tweet analysis

Alam et al. [20] proposed Image4Act, a deep neural network image classification framework, leveraging VGG16 architecture, which achieved an accuracy rate of 67%. In study [21], the utilization of VGG16 and perceptual hashing was explored for image classification pertaining to damage assessment. The findings emphasized the potential of social media photos to guide emergency response efforts. Chaudhuri and Bose [22] introduced a CNN model with an accuracy of 83.2% to identify debris along with human body parts.

A lightweight CNN featuring two categorization heads, proposed by Valdez and Godmalin [23], was tailored for discerning both the type and intensity of natural disasters. Employing MobileNetV3 and FFN, the study collected data spanning "wildfire," "flood," "earthquake," and "volcanic eruption" at varying severity levels, achieving a remarkable 96.8% accuracy in disaster type classification and 93.2% accuracy in intensity level determination.

For aerial image categorization captured by unmanned aerial vehicles (UAV), Kyrkou and Theocharides [24] formulated a CNN model with residual connections. Aerial shots encompassing "fire," "flooding," "wrecked buildings," and "road accidents" were included, yielding an accuracy rate of 90.1%.

2.3 Multimodal based classification in disaster tweet analysis

In study of Rizk et al. [25], a two-stage multimodal framework tailored for energy-constrained devices was devised to manage both text and image data from Twitter. In the first level, classifiers processed images and text, and in the second level, their decisions were combined, resulting in a 92.43% accuracy rate. Mouzannar et al. [26] introduced a multimodal deep learning (DL) model for damage detection, utilizing CNN-based neural networks for text and pretrained inception models for images. The combined features underwent classification by FFN, yielding an accuracy of 92.62%.

A multimodal approach with four distinct classifiers to filter Hurricane Irma-related tweets containing geospatial information was presented by Mohanty et al. [27]. These classifiers encompassed an image classifier, a user authenticity classifier, and a text classifier. The integrated outputs of these models led to the identification of social media posts useful during calamities once a predefined threshold score was reached. Kumar et al. [28] proposed an alternative multimodal method for categorizing disaster-related informative content. By combining the outputs of LSTM and VGG16, the feature vector was fed into an FFN, yielding F1-scores ranging from 0.61 to 0.92 across multiple datasets. Addressing both text and image aspects of tweets, Madichetty et al. [29] introduced a multimodal approach based on BERT and Densenet.

In the study [30], an image was classified using VGG16, while text was classified through word2vec and CNN in a deep learning neural network. The ultimate classification was conveyed to an FFN as a merged image feature vector and text feature vector, achieving an accuracy of 78.4%.

The existing multimodal fusion methods face limitations in dealing with intricate and high-dimensional multimodal data. This is attributed to:

- The current systems considering inputs from only one modality, learning patterns within that modality exclusively.
- Lack of capacity in existing systems to prioritize feature importance and establish connections across diverse input modalities, without assigning varying values to specific characteristics.

To address this challenge, a unified theory is required to steer the development of feature-fusion-centered classification techniques that can unveil meaningful latent associations across various modalities. Such a theory should assign differential emphasis to characteristics based on their relative significance [31].

3. PRELIMINARIES

This segment will provide an overview of the key aspects of the proposed system. The recommended framework comprises several integral components: a module for preprocessing text, a text model employing the Robustly Optimized BERT Pretraining Approach, an image model utilizing the Vision Transformer, a bidirectional long short-term memory (BiLSTM), and an attention module.

3.1 Preparation of dataset

The data represented in textual form lacks structure, and through the organization and standardization of raw data, pre-processing enhances the capability of the text processing system to effectively handle the data [32]. A typical social media tweet, often limited to 140 characters, contains repetitions of the same information. Pre-processing aims to mitigate the impact of these issues. Our extensive testing has revealed that pre-processed text leads to improved performance [33]. Common techniques employed in text pre-processing encompass actions such as "Expanding contractions", "Converting letters to lowercase", "Eliminating punctuation", "Omitting stop words", "Excluding URLs", "Stemming and lemmatization", and "Trimming extra spaces".

3.2 How word embeddings work

A "word embedding" is a real-valued vector that represents text through distributed learning. It portrays words with similar meanings in a largely equivalent manner. This incorporation of deep learning has marked a significant advancement in the field of Natural Language Processing.

A pivotal moment in natural language processing (NLP) transpired in 2018 when Google Brain introduced BERT (Bi-directional Encoder Representations from Transformers), utilizing sources like Book Corpus and Wikipedia for pretraining [34]. Built upon the transformer architecture, BERT transformed conventional encoder-decoder systems. Subsequent refinements to BERT's performance and training speed include XLNet, RoBERTa, and DistilBERT. In over 20 natural language processing tasks, RoBERTa outperforms BERT across numerous benchmark datasets.

The concept of fine-tuning can be likened to a form of transfer learning. Transfer learning, as defined by Bengio et al. [35], involves applying knowledge acquired in one context to solve problems in another.

A pretrained model designed for one task can be adapted for a different task with minor adjustments. The learned weights from the pretrained model are used to train a new model for the desired task. Fine-tuning proves particularly advantageous when the labeled dataset is limited, reducing the risk of overfitting and expediting model convergence while conserving computational resources and time.

Tripathy et al. [36] undertook fine-tuning of the ALBERT language model to detect instances of cyberbullying within social media data, achieving an impressive F1 score of 95%. In a similar vein, Sindhura et al. [37] conducted sentiment analysis on mobile app reviews by Indonesian users, employing both a customized BERT model and a pretrained model. Their use of the pretrained model resulted in state-of-the-art performance.

In the realm of grammatical correction across multiple languages, Pająk and Pająk [38] and Reddy et al. [39] devised an NLP system. Despite working with a smaller dataset and limited processing resources, their findings highlighted the effectiveness of fine-tuning in producing superior results.

For our endeavor, we utilized a RoBERTa model that had been previously trained and then adapted for our specific task-oriented dataset. RoBERTa encompasses several key attributes:

•Generating context-aware word embeddings.

•Implementing the bidirectional transformer concept, which enables the utilization of historical and future contexts.

•Employing a byte-pair encoding technique for tokenization, allowing for the comprehension of uncommon terms and tokens that may not be prevalent in the language's lexicon.

3.3 Utilizing bidirectional long short-term memory (Bi-LSTM) for enhanced disaster tweet classification

In the context of disaster tweet classification, the usage of Bidirectional Long Short-Term Memory (Bi-LSTM) is pivotal. Bi-LSTM is a recurrent neural network (RNN) architecture tailored for processing sequential data, such as text. For the purpose of classifying disaster-related tweets, Bi-LSTM is harnessed to capture the inherent sequential patterns and context embedded within the textual content of the tweets.

The equations for the forward and backward passes of a Bi-LSTM cell are described from Eqs. (1)-(12).

Forward Pass:

$$\text{Forget Gate: } f_t = \sigma\left(W_{f1} \cdot [h_{t-1}, x_t] + b_{f1}\right) \tag{1}$$

$$\text{Input Gate: } i_t = \sigma(W_{i1} \cdot [h_{t-1}, x_t] + b_{i1}) \tag{2}$$

$$\text{Candidate Cell State: } \tilde{C}_t = \tanh\,(W_{C1} \cdot [h_{t-1}, x_t] + b_{C1}) \tag{3}$$

$$\text{Cell State: } C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{4}$$

$$\text{Output Gate: } o_t = \sigma(W_{o1} \cdot [h_{t-1}, x_t] + b_{o1}) \tag{5}$$

$$\text{Hidden State: } h_t = o_t \cdot \tanh\,(C_t) \tag{6}$$

Backward Pass:

$$\text{Forget Gate: } f_{t'} = \sigma\left(W_{bf1} \cdot [h_{t'+1}, x_{t'}] + b_{bf1}\right) \tag{7}$$

$$\text{Input Gate: } i_{t'} = \sigma(W_{bi1} \cdot [h_{t'+1}, x_{t'}] + b_{bi1}) \tag{8}$$

$$\text{Candidate Cell State: } \tilde{C}_{t'} = \tanh\,(W_{bC1} \cdot [h_{t'+1}, x_{t'}] + b_{bC1}) \tag{9}$$

$$\text{Cell State: } C_{t'} = f_{t'} \cdot C_{t'+1} + i_{t'} \cdot \tilde{C}_{t'} \tag{10}$$

$$\text{Output Gate: } o_{t'} = \sigma(W_{bo1} \cdot [h_{t'+1}, x_{t'}] + b_{bo1}) \tag{11}$$

$$\text{Hidden State: } h_{t'} = o_{t'} \cdot \tanh\,(C_{t'}) \tag{12}$$

In these equations, the notations x_t represent the input at time step t, h_t signifies the hidden state at time step t, C_t embodies the cell state at time step t, and σ denotes the sigmoid activation function. The weights W_{f1}, W_{i1}, W_{C1}, W_{o1}, W_{bf1}, W_{bi1}, W_{bC1}, W_{bo1} coupled with bias terms b_{f1}, b_{i1}, b_{c1}, b_{o1}, b_{bf1}, b_{bi1}, b_{bC1}, b_{bo1} are fine-tuned during training.

In the context of classifying disaster-related tweets, the Bi-LSTM mechanism is applied to the tokenized and embedded textual data of the tweets. Subsequently, the outputs of both the forward and backward passes are concatenated to yield a comprehensive grasp of the sequential nuances present in the text. This enriched representation then serves as input for subsequent layers within the classification model, greatly aiding the accurate categorization of tweets into informative or non-informative classes for efficient disaster management.

3.4 Utilizing attention mechanism for enhanced disaster tweet classification

In the context of disaster tweet classification, integrating the Bahdanau attention mechanism has proven to be a potent approach for improving model performance. The Bahdanau attention mechanism enables the model to dynamically assign varying degrees of importance to different elements within the input sequence, allowing the model to focus on the most relevant information for accurate classification.

The Bahdanau attention mechanism involves a series of learned parameters and computations to determine attention weights for individual elements in the input sequence. These attention weights are then utilized to compute a context vector that captures the most salient information.

Mathematically, the Bahdanau attention mechanism [40] can be described as follows:

Given an input sequence of embeddings X=[x_1, x_2, ..., x_n], where x_i represents the embedding of the i-th word in the sequence, and a context vector C, the attention scores e_i for each element can be calculated using a compatibility function:

$$e_i=f(x_i,C) \tag{13}$$

The attention scores are then transformed into attention weights using a softmax function:

$$\alpha_i=\exp(e_i)/\Sigma_j\exp(e_j) \tag{14}$$

The context vector C is obtained by computing the weighted sum of the input sequence embeddings using the calculated attention weights:

$$C=\Sigma_i\alpha_i\times x_i \tag{15}$$

Finally, the context vector C is combined with the hidden state of the model to generate the ultimate output.

By incorporating the Bahdanau attention mechanism, the model gains the ability to dynamically allocate attention to different parts of the input sequence, enhancing its focus on critical components for disaster tweet classification. This adaptability results in improved accuracy and more robust handling of intricate and informative tweets.

The integration of Bahdanau attention mechanisms into disaster tweet classification models enhances the model's capability to capture contextual nuances within the text, leading to elevated accuracy and effectiveness in the classification process.

An attentional idea is instantiated as a feed forward neural network (FFN). In order to determine which encoded input vectors are most important, the FFN assigns attention scores, with the higher attention scores going to the most relevant input vectors. The attention score describes how important each input is to accomplishing the job. A softmax operation is then applied to normalize this score. Subsequently, the input data's context vector is formulated by merging the attention-weighted input vectors. The following Eq. (16) and Eq. (17). show the entire process.

$$\alpha_t=\frac{exp(v.h_t)}{\sum exp(v.h_t)} \tag{16}$$

$$S=\sum_t \alpha_t\, h_t \tag{17}$$

where, h_t is Hidden vector, α_t is Attention weight and S is Attention-weighted context vector.

4. PROPOSED SYSTEM

4.1 Overview

In the context of disaster scenarios, where the vast volume of conversational data on social media holds crucial situational insights, we present a novel neural network architecture aimed at enhancing accuracy. This architecture takes into consideration both textual content and accompanying images, offering the potential for disaster relief teams to swiftly streamline recovery efforts post-crisis. The intricate system design is depicted in Figure 2, showcasing a comprehensive blueprint of our proposed approach. Demonstrating superior performance compared to existing models, our proposed architecture stands as a remarkable advancement.

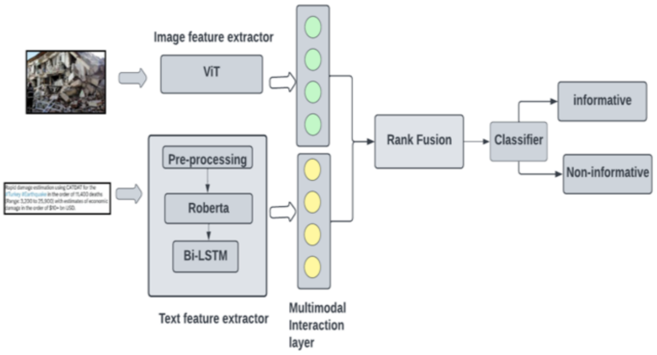


Figure 2. Proposed system architecture

Central to our proposed approach are several pivotal components:

•Textual Feature Extraction Module: This module is responsible for extracting pertinent information from tweet text.

•Image Feature Extraction Module: Extracting features from associated images is the prime function of this module.

•Fusion Module: Combining features from diverse modules, this module employs an early fusion technique.

•Classification Module: Final classification tasks are undertaken by a feed-forward neural network (FFN) as the concluding step.

4.2 Feature extraction modules

Text feature extraction using both RoBERTa and Bi-LSTM involves processing the input text data to obtain meaningful representations that capture the semantic context and sequential patterns within the text. Here's a brief overview of how these two methods work:

RoBERTa for Text Feature Extraction

RoBERTa is a transformer-based language model designed for natural language understanding tasks. It processes text input by dividing it into tokens and then generating contextualized embeddings for each token. The embeddings capture the semantic meaning of the words within the context of the entire text. The process involves the following steps:

-Tokenization: The input text is split into individual tokens (words or subwords).

-Embedding: Each token is converted into a vector representation (embedding).

-Transformer Encoding: The embeddings are passed through multiple transformer layers, which capture relationships between tokens and generate contextualized embeddings.

-Pooling: The embeddings from the last layer can be aggregated using pooling techniques, such as mean pooling or max pooling, to obtain an overall representation of the text.

Bi-LSTM for Text Feature Extraction

Bi-LSTM (Bidirectional Long Short-Term Memory) is a type of recurrent neural network designed to capture sequential patterns in text. It processes the text input token by token, considering both past and future contexts. The process involves the following steps:

-Tokenization: Similar to RoBERTa, the input text is tokenized into individual units.

-Embedding: Each token is converted into a vector representation using pre-trained word embeddings.

-Bi-LSTM Processing: The embeddings are passed through a bidirectional LSTM network. For each token, the LSTM cells process the sequence in both forward and backward directions, capturing contextual information.

-Pooling or Aggregation: Similar to RoBERTa, you can use pooling techniques to aggregate the LSTM outputs to obtain a single representation of the entire text.

Both RoBERTa and Bi-LSTM provide different approaches to text feature extraction. RoBERTa excels in capturing contextual semantics, while Bi-LSTM is effective at capturing sequential dependencies. By using both methods, you can combine their strengths and create a more comprehensive text representation. In your proposed approach, these extracted text features can be further combined with image features to enhance the performance of your disaster tweet classification system.

Image Feature Extraction using VIT

VIT (Vision Transformer) [41] is a state-of-the-art model for image classification. It divides the input image into fixed-size patches and linearly embeds them into vectors. These patch embeddings, along with positional embeddings, are fed into a transformer encoder. The transformer captures the spatial relationships between different patches, allowing it to extract image features. The final layer's output is used as the image representation, which encodes the salient visual information present in the image input.

In the proposed disaster tweet classification approach, the extracted text features from RoBERTa and LSTM, as well as the image feature from VIT, are combined to form a comprehensive representation of the tweet. This combined representation leverages both textual and visual information to enhance the classification accuracy [42, 43], enabling the system to better distinguish between informative and non-informative tweets during disasters.

4.3 Rank fusion module

Rank fusion is a technique used to combine the ranking orders of multiple classifiers or models in order to produce a consolidated ranking order that captures the strengths of individual rankings. It is commonly applied in scenarios where multiple models provide different perspectives on the same data, and the goal is to create a more accurate and robust ranking or classification outcome.

Rank the predicted class probabilities or confidence scores for each tweet from both text and image classifiers. Lower-ranked predictions are considered stronger and more informative. Combine the rankings from text and image predictions using a fusion technique. Common approaches include weighted sum, weighted average, or even more advanced methods like Borda count or logistic regression. Assign appropriate weights to the rankings from text and image classifiers. These weights can be determined based on validation performance or domain knowledge. Calculate the combined rank for each tweet using the fused rankings. This can involve adding or averaging the ranks from text and image modalities, depending on the chosen fusion approach. Apply a threshold to the combined rank to determine the final classification. If the combined rank is below the threshold, classify the tweet as informative; otherwise, classify it as non-informative.

5. EXPERIMENTAL SETUP

In this section, we elaborate on the experimental configuration and environment that we utilized during the execution phase. We ensured that both training and testing procedures were carried out on the same dataset to ensure the practical applicability of the proposed system.Our focus centered on the analysis of multimodal data, and to accomplish this, we leveraged the CrisisMMD dataset as outlined in study [3]. This dataset has been extensively employed by numerous advanced models [13, 28, 29], showcasing its significant utility.

The CrisisMMD dataset comprises text and corresponding images extracted from Twitter, covering seven distinct natural disasters that occurred in 2017, including events like fires, earthquakes, and floods. This dataset exhibits a diverse range of linguistic variations and semantic contexts. Each tweet within the dataset is assigned labels indicating whether it is informative or non-informative.

In order to establish a comprehensive and insightful dataset, we undertook the task of data collection, followed by thorough data cleansing and augmentation processes. All seven of these curated datasets were then employed for both training and testing the proposed model. A breakdown of the dataset distribution is presented in Table 1.

Table 1. Datasets related to disasters

Disaster Name	Informative Text	Informative Image	Non-Informative Text	Non-Informative Image
Hurricane Irma	3544	2208	960	2296
Hurricane Harvey	3332	2457	1102	1977
Hurricane Maria	2842	2231	1714	2325
California Wildfires	1253	985	336	604
Mexico Earthquake	1031	841	349	539
Iraq Iran Earthquake	493	400	104	197
Sri Lanka Floods	367	252	655	770

The development of our models was conducted within the TensorFlow framework using the Python programming language. In particular, we fine-tuned and harnessed the RoBERTa and ViT models. Our experimentation outcomes played a pivotal role in guiding the selection of optimal feature extractors for both text and image inputs. It's worth highlighting that the RoBERTa and ViT models demonstrated remarkable performance levels throughout our experiments.

In this study different combinations of fusion strategies were used.

M1: "additive fusion with RoBERTa & ViT"
M2: "concatenative fusion with RoBERTa & ViT"
M3: "averaged fusion with RoBERTa & ViT"
M4: "multiplicative fusion with RoBERTa & ViT"
M5: "Rank fusion"

We tried out a few distinct multimodal fusion methods and compared the results to the proposed method and certain already-established platforms. Classifications were made both within across the domains.

Fusion by addition: The addition operation combines many inputs into one. This works well for applications where the total input numbers don't make a huge difference.

Fusion with concatenation: The various inputs are combined by being concatenated together. An advantage of concatenation is that the inputs are not changed or are only changed slightly, hence this is an argument in its favour. This ensures that the inputs maintain their native appearance and behaviour.

Fusion with average: The fused vector is calculated by averaging the inputs.

Fusion by multiplication: When several inputs are multiplied together, a unified output is produced. Acquiring knowledge of how various sensory modalities interact is a strong suit of multiplicative fusion.

Fusion by Rank:

Rank fusion is a technique used to combine the ranking orders of multiple classifiers or models in order to produce a consolidated ranking order that captures the strengths of individual rankings.

6. DISCUSSIONS

Using CrisisMMD, a standard multimodal dataset, we evaluated our approach. By combining image and text analysis, we have conducted extensive studies. After pooling data from each of the seven crises covered by the CrisisMMD, we were able to use it as a unified set for classification purposes. Tables 2-8, show the classification experiment results. The best results we found in our experiments for each performance parameter are highlighted in the tables. The results showed that the proposed system significantly outperformed the state-of-the-art systems in all of the aforementioned evaluation metrics.

The proposed approach consistently exhibits superior performance compared to alternative methods in all evaluation scenarios. This technique introduces a multimodal data fusion framework for classifying tweets, where both the textual content of the tweet and the associated image are utilized as input.

Through the collaborative utilization of text and image inputs, their individual strengths are mutually enhanced. Each form of information is assessed in conjunction with the other, allowing for a deeper understanding of the contextual background. In our approach, text feature extraction was accomplished using the RoBERTa model, while image features were extracted using the ViT model.

Table 2. Outcomes of multi-modal fusion techniques applied on the dataset California Wildfires

Model	Accuracy	Precision	Recall	F1-Measure
M1	94.16	94	95	95
M2	94.26	94	95	95
M3	96.53	97	98	97
M4	96.13	96	97	97
M5	97.04	96.05	96.72	96

Table 3. Outcomes of multi-modal fusion techniques applied to the dataset hurricane

Model	Accuracy	Precision	Recall	F1-Measure
M1	92.91	94	95	94
M2	95.01	96	96	96
M3	94	95	96	94
M4	96	96	97	96
M5	**96.49**	**97.45**	**97.42**	**97.9**

Table 4. Outcomes of multi-modal fusion techniques applied on the dataset Hurricane Irma

Model	Accuracy	Precision	Recall	F1-Measure
M1	92.32	92	92	92
M2	92.57	93	92	92
M3	92.11	92	91	92
M4	98.24	97	97	98
M5	97.49	98.25	98.25	98.25

Table 5. Outcomes of multi-modal fusion techniques applied on the dataset Hurricane Maria

Model	Accuracy	Precision	Recall	F1-Measure
M1	92.72	92	92	92
M2	92.85	93.78	92	96
M3	92.91	92	91	92
M4	98.84	98	98	98
M5	97.89	98.15	98.15	98.05

Table 6. Outcomes of multi-modal fusion techniques applied on the dataset Iraq-Iran Earthquake

Model	Accuracy	Precision	Recall	F1-Measure
M1	93.94	94	94	94
M2	96.46	96	96	96
M3	94.95	95	95	95
M4	97.98	98	98	98
M5	98.33	98	98	98

Table 7. Outcomes of multi-modal fusion techniques applied on the dataset Mexico Earthquake

Model	Accuracy	Precision	Recall	F1-Measure
M1	92.76	93	93	93
M2	92.52	93.14	93	93
M3	93.22	93	93	93
M4	97.9	98	98	98
M5	96.99	98.29	98.29	98.29

Table 8. Outcomes of multi-modal fusion techniques applied on the dataset Sri Lanka Floods

Model	Accuracy	Precision	Recall	F1-Measure
M1	96.84	97	97	97
M2	94.47	94	94	94
M3	97.23	96	96	97
M4	97.23	96	96	97
M5	97.45	96.32	96.32	97.32

Various fusion methods, including addition, concatenation, averaging, and multiplication, were benchmarked against the proposed system. The concatenation-based fusion method fails to explore the interactions between diverse input modalities. Additive and averaged fusion methods lack comprehensive examination of the collective value of features. In contrast, rank fusion effectively unveils latent correlations between feature vectors of different modalities, thereby capturing intricate relationships across various input types.

The additive method is employed to generate potential feature mixtures, while the multiplicative method subsequently selects impactful modality combinations. In tandem with this fusion strategy, we integrated Bi-LSTM and an attention mechanism. Bi-LSTM facilitates bidirectional context analysis, allowing for a deeper understanding of contextual nuances. The attention layer, when combined with Bi-LSTM, prioritizes highly contextualized, fine-grained features, thus optimizing their relevance for downstream tasks. This comprehensive architecture significantly elevates performance outcomes.

Collectively, the proposed system demonstrates consistent success across diverse datasets, showcasing its scalability potential. Its stability across various domains underscores its versatility. After careful deliberation, it is evident that our proposed approach provides a pragmatic solution to the challenge of tweet classification in times of crisis.

7. CONCLUSION

We introduced a novel neural network architecture tailored for disaster management systems, aimed at classifying pertinent information within user-generated social media posts. This innovative system leverages both textual and visual content to formulate its classifications. By seamlessly integrating natural language processing, image analysis, and advanced deep learning techniques, the potential for achieving elevated performance outcomes is effectively harnessed.

Within the proposed system, we have ingeniously crafted an rank fusion mechanism that bridges the gap between different modalities. Employing the text feature extraction prowess of RoBERTa alongside the image feature extraction capabilities of ViT, a comprehensive approach is forged. The experimentation phase entails the utilization of seven diverse multimodal disaster datasets encompassing categories such as "hurricanes," "earthquakes," "wildfires," and "floods." The robustness of the system is scrutinized across these datasets, and the results are statistically promising, underscoring its robust reliability. Importantly, the observed enhancements in accuracy over baseline models are notably substantial.

REFERENCES

[1] Palen, L., Anderson, K.M., Mark, G., Martin, J., Sicker, D., Palmer, M., Grunwald, D. (2010). A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. ACM-BCS Visions of Computer Science, 2010. https://doi.org/10.14236/ewic/VOCS2010.8

[2] Martínez-Rojas, M., del Carmen Pardo-Ferreira, M., Rubio-Romero, J.C. (2018). Twitter as a tool for the management and analysis of emergency situations: A systematic literature review. International Journal of Information Management, 43: 196-208. https://doi.org/10.1016/j.ijinfomgt.2018.07.008

[3] Alam, F., Ofli, F., Imran, M. (2018). Crisismmd: Multimodal twitter datasets from natural disasters. Twelfth International AAAI Conference on Web and Social Media, 12(1): 465-473. https://doi.org/10.1609/icwsm.v12i1.14983

[4] Shah, R., Zimmermann, R. (2017). Multimodal analysis of user-generated multimedia content. Cham: Springer International Publishing.

[5] Shah, R.R., Yu, Y., Zimmermann, R. (2014). Advisor: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando Florida, USA, pp. 607-616. https://doi.org/10.1145/2647868.2654919

[6] Shah, R.R., Mahata, D., Choudhary, V., Bajpai, R. (2018). Multimodal semantics and affective computing from multimedia content. In Intelligent Multidimensional Data and Image Processing, pp. 359-382. https://doi.org/10.4018/978-1-5225-5246-8.ch014

[7] Yu, Y., Tang, S., Aizawa, K., Aizawa, A. (2018). Category-based deep CCA for fine-grained venue discovery from multimodal data. IEEE Transactions on Neural Networks and Learning Systems, 30(4): 1250-1258. https://doi.org/10.1109/TNNLS.2018.2856253

[8] Yu, Y., Tang, S., Raposo, F., Chen, L. (2019). Deep cross-modal correlation learning for audio and lyrics in music retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15(1): 20. https://doi.org/10.1145/3281746

[9] Sirisha, U., Bolem, S.C. (2022). Aspect based sentiment & emotion analysis with ROBERTa, LSTM. International Journal of Advanced Computer Science and Applications, 13(11): 766-774. https://doi.org/10.14569/IJACSA.2022.0131189

[10] Gunes, H., Piccardi, M. (2008). Automatic temporal segment detection and affect recognition from face and body display. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(1): 64-84. https://doi.org/10.1109/TSMCB.2008.927269

[11] Sirisha, U., Sai Chandana, B. (2022). Semantic interdisciplinary evaluation of image captioning models. Cogent Engineering, 9(1): 2104333. https://doi.org/10.1080/23311916.2022.2104333

[12] Sirisha, U., Chandana, B.S. (2023). Privacy preserving image encryption with optimal deep transfer learning based accident severity classification model. Sensors, 23(1): 519. https://doi.org/10.3390/s23010519

[13] Madichetty, S., Sridevi, M. (2021). A novel method for identifying the damage assessment tweets during disaster. Future Generation Computer Systems, 116: 440-454. https://doi.org/10.1016/j.future.2020.10.037

[14] Madichetty, S. (2021). A stacked convolutional neural network for detecting the resource tweets during a

disaster. Multimedia Tools and Applications, 80: 3927-3949. https://doi.org/10.1007/s11042-020-09873-8

[15] Snyder, L.S., Lin, Y.S., Karimzadeh, M., Goldwasser, D., Ebert, D.S. (2019). Interactive learning for identifying relevant tweets to support real-time situational awareness. IEEE Transactions on Visualization and Computer Graphics, 26(1): 558-568. https://doi.org/10.1109/TVCG.2019.2934614

[16] Madichetty, S., Sridevi, M. (2020). Identification of medical resource tweets using majority voting-based ensemble during disaster. Social Network Analysis and Mining, 10: 66. https://doi.org/10.1007/s13278-020-00679-y

[17] Zahra, K., Imran, M., Ostermann, F.O. (2020). Automatic identification of eyewitness messages on twitter during disasters. Information Processing & Management, 57(1): 102107. https://doi.org/10.1016/j.ipm.2019.102107

[18] Ghafarian, S.H., Yazdi, H.S. (2020). Identifying crisis-related informative tweets using learning on distributions. Information Processing & Management, 57(2): 102145. https://doi.org/10.1016/j.ipm.2019.102145

[19] Kejriwal, M., Zhou, P. (2020). On detecting urgency in short crisis messages using minimal supervision and transfer learning. Social Network Analysis and Mining, 10(1): 58. https://doi.org/10.1007/s13278-020-00670-7

[20] Alam, F., Imran, M., Ofli, F. (2017). Image4act: Online social media image processing for disaster response. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, pp. 601-604. https://doi.org/10.1145/3110025.3110164

[21] Alam, F., Ofli, F., Imran, M. (2018). Processing social media images by combining human and machine computing during crises. International Journal of Human–Computer Interaction, 34(4): 311-327. https://doi.org/10.1080/10447318.2018.1427831

[22] Chaudhuri, N., Bose, I. (2019). Application of image analytics for disaster response in smart cities. Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 3036-3045. https://doi.org/10.24251/HICSS.2019.367

[23] Valdez, D.B., Godmalin, R.A.G. (2021). A deep learning approach of recognizing natural disasters on images using convolutional neural network and transfer learning. In Proceedings of the International Conference on Artificial Intelligence and its Applications, Virtual Event, pp. 1-7. https://doi.org/10.1145/3487923.3487927

[24] Kyrkou, C., Theocharides, T. (2019). Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Beach, CA, USA, pp. 517-525. https://doi.org/10.1109/CVPRW.2019.00077

[25] Rizk, Y., Jomaa, H. S., Awad, M., Castillo, C. (2019). A computationally efficient multi-modal classification approach of disaster-related Twitter images. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Limassol, Cyprus, pp. 2050-2059. https://doi.org/10.1145/3297280.3297481

[26] Mouzannar, H., Rizk, Y., Awad, M. (2018). Damage identification in social media posts using multimodal deep learning. In Proceedings of the 15th ISCRAM Conference – Rochester, NY, USA.

[27] Mohanty, S.D., Biggers, B., Sayedahmed, S., Pourebrahim, N., Goldstein, E.B., Bunch, R., Chi, G., Sadri, F., McCoy, T.P., Cosby, A. (2021). A multi-modal approach towards mining social media data during natural disasters-A case study of Hurricane Irma. International Journal of Disaster Risk Reduction, 54: 102032. https://doi.org/10.1016/j.ijdrr.2020.102032

[28] Kumar, A., Singh, J. P., Dwivedi, Y.K., Rana, N.P. (2020). A deep multi-modal neural network for informative Twitter content classification during emergencies. Annals of Operations Research, 319: 791-822. https://doi.org/10.1007/s10479-020-03514-x

[29] Madichetty, S., Muthukumarasamy, S., Jayadev, P. (2021). Multi-modal classification of Twitter data during disasters for humanitarian response. Journal of Ambient Intelligence and Humanized Computing, 12: 10223-10237. https://doi.org/10.1007/s12652-020-02791-5

[30] Ofli, F., Alam, F., Imran, M. (2020). Analysis of social media data using multimodal deep learning for disaster response. arXiv preprint arXiv:2004.11838. https://doi.org/10.48550/arXiv.2004.11838

[31] Sirisha, U., Praveen, S.P., Srinivasu, P.N., Barsocchi, P., Bhoi, A.K. (2023). Statistical analysis of design aspects of various YOLO-based deep learning models for object detection. International Journal of Computational Intelligence Systems, 16(1): 126. https://doi.org/10.1007/s44196-023-00302-w

[32] Marrapu, B.V., Raju, K.Y.N., Chowdary, M.J., Vempati, H., Praveen, S.P. (2022). Automating the creation of machine learning algorithms using basic math. In 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 866-871. https://doi.org/10.1109/ICSSIT53264.2022.9716270

[33] Singh, T., Kumari, M. (2016). Role of text pre-processing in twitter sentiment analysis. Procedia Computer Science, 89: 549-554. https://doi.org/10.1016/j.procs.2016.06.095

[34] Devlin, J., Chang, M. W., Lee, K., Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

[35] Bengio, Y., Courville, A., Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8): 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

[36] Tripathy, J.K., Chakkaravarthy, S.S., Satapathy, S.C., Sahoo, M., Vaidehi, V. (2022). ALBERT-based fine-tuning model for cyberbullying analysis. Multimedia Systems, 28(6): 1941-1949. https://doi.org/10.1007/s00530-020-00690-5

[37] Sindhura, S., Phani Praveen, S., Madhuri, A., Swapna, D. (2022). Different feature selection methods performance analysis for intrusion detection. Smart Intelligent Computing and Applications, 2: 523-531. https://doi.org/10.1007/978-981-16-9705-0_51

[38] Pająk, K., Pająk, D. (2022). Multilingual fine-tuning for grammatical error correction. Expert Systems with Applications, 200: 116948.

https://doi.org/10.1016/j.eswa.2022.116948

[39] Reddy, A.S., Praveen, S.P., Ramudu, G.B., Anish, A.B., Mahadev, A., Swapna, D. (2023). A network monitoring model based on convolutional neural networks for unbalanced network activity. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, pp. 1267-1274. https://doi.org/10.1109/ICSSIT55814.2023.10060879

[40] Sindhura, S., Praveen, S.P., Safali, M.A., Rao, N. (2021). Sentiment analysis for product reviews based on weakly-supervised deep embedding. In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, pp. 999-1004. https://doi.org/10.1109/ICIRCA51532.2021.9544985

[41] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929

[42] Bahdanau, D., Cho, K., Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. https://doi.org/10.48550/arXiv.1409.0473

[43] Srinivasu, P.N., JayaLakshmi, G., Jhaveri, R.H., Praveen, S.P. (2022). Ambient assistive living for monitoring the physical activity of diabetic adults through body area networks. Mobile Information Systems, 2022: 3169927. https://doi.org/10.1155/2022/3169927