Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-025-10314-7

RESEARCH ARTICLE-COMPUTER ENGINEERING AND COMPUTER SCIENCE

Analysis of Multimodal Social Media Data Utilizing VIT Base 16 and GPT-2 for Disaster Response

Shilpa Gite[1,2] · Shruti Patil[1,2] · Biswajeet Pradhan[3] · Madhuri Yadav[1] · Sneha Basak[1] · Arundarasi Rajendra[1] · Abdullah Alamri[4] · Kaustubh Raykar[1] · Ketan Kotecha[1,2]

Received: 3 May 2024 / Accepted: 5 May 2025

Abstract
Multimedia systems, such as social media platforms, play a crucial role in disseminating vital information during calamities. This information is shared in various formats such as images, text, videos, audio, etc. Therefore, it becomes important to have a system that can identify multimodal data to classify relevant information. This paper proposes a new age classification method for multimodal data using advanced and improved transformer models, such as Vision Transformer and Generative Pre-trained Transformer 2, for image and text classification, respectively. These models were combined using an ensemble model (Random Forest Classifier), achieving an accuracy of 84.66% on the multimodal data. Furthermore, the proposed model demonstrates higher prediction accuracy compared to traditional Convolutional Neural Network (CNN) models which have an accuracy of 71.43%, exceeding it by 13.23%. A comparison with convolutional models is conducted to underscore the advantages of transformer models and to substantiate the necessity of the experiment. Our proposed classification model using Vision Transformer and GPT-2, along with an ensemble model, can be replicated by researchers in disaster management, humanitarian aid organizations, and social media platforms looking to filter and prioritize information during emergencies.

Keywords Multimodal data · Vision Transformer (ViT) · Transformer-based language model · Generative pre-trained transformer (GPT-2) · Disaster management

1 Introduction

In the actual world, data are generally presented in various formats. Seeing objects, hearing sounds, smelling scents, and so forth are common experiences as we interact with the environment. Different instruments, measuring techniques, and other sources are employed to collect information about an event or a system of interest. Due to the diverse properties of natural processes and habitats, it is uncommon for a single acquisition approach to offer a comprehensive understanding of them. The growing availability of multiple datasets containing information about the same system, gathered through various acquisition methods, introduces new degrees of freedom [1].

Modality refers to the way something happens or is experienced. Data available across a wide range of such modalities is known as multimodal data. Usually, when a user interacts with a system, multimodal data is generated in massive amounts. Finding a proper way to analyze and interpret such data is something that users still struggle with [2]. One such sector where understanding this data can aid is an analysis of events using social media. It is something that researchers have worked around a lot for the past few years [2]. People have proposed numerous solutions and approaches for analyzing these events [3]. Natural disasters serve as an example of an event in which a significant portion of this event-related data is collected via a network of peers [4].

✉ Shilpa Gite
shilpa.gite@sitpune.edu.in

✉ Biswajeet Pradhan
biswajeet.pradhan@uts.edu.au

1 Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India

2 Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, Maharashtra 412115, India

3 Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia

4 Department of Geology and Geophysics, College of Science, King Saud University, Riyadh, Saudi Arabia

Published online: 31 May 2025

Data sharing has rapidly expanded due to the advancement of social media networks. The data format on these networks has evolved from users solely being able to share texts to a more advanced where users can now share more than just text. We foresee a massive surge in multimodal posts since it is observed that posts with photos increase interaction while tweets containing videos increase engagement [5]. In the context of natural disaster-related posts, while such information is essential to people providing humanitarian assistance, these multimodal posts being shared have resulted in a slew of other issues. One of the most fundamental issues is that non-informative posts are mixed with relevant ones [4]. Hence, the classification of social media content is essential.

The majority of earlier research on analyzing social media content for crisis-related situations focused mainly on textual data, with little attention paid to images uploaded on social media. Several previous studies have shown that images put up by users on social media after a crisis can help relief organizations in a variety of ways.

For instance, [6] determined how severely the resources were damaged using images uploaded on Twitter, while [7] went a step further and focused on identifying the damaged features.

In this study, we aim to use both textual and visual contents of Twitter data to learn whether a tweet is informative or not. Given the limitations of unimodal techniques, various studies have been conducted to analyze multimodal data for social media research. These studies have been shown to enhance classification accuracy [5]. However, most of these methods are complex and rely on cutting-edge models such as convolutional neural networks and ResNet50. In response to these findings, this study provides a unique framework for multimodal classification.

The novelty of this paper lies in its innovative approach to multimodal data classification for disaster response using advanced transformer models. Unlike traditional methods that depend heavily on Convolutional Neural Networks (CNNs) and simpler language models, our approach integrates Vision Transformer (ViT Base 16) for image classification and Generative Pre-trained Transformer 2 (GPT-2) for text classification. By combining these models through an ensemble Random Forest Classifier, we achieve a significant improvement in accuracy, outperforming existing CNN-based models by 13.23%. ViT was selected for its ability to capture long-range dependencies in images effectively, while GPT-2 was chosen for its superior contextual understanding of text. This combination ensures robust feature extraction from both modalities. This dual-transformer approach not only harnesses the powerful capabilities of ViT and GPT-2 but also effectively leverages their strengths in processing visual and textual data, respectively. Our study demonstrates the efficacy of this combination in the context of disaster-related social media data, providing a robust and scalable solution for rapid and accurate information dissemination during emergencies.

2 Related Work

This section gives an overview of the literature review of relevant research papers on multimodal classification.

In their work, [5] proposed multimodal classification for analyzing social media data, which utilizes multimodal data for emotion classification through neural network models to merge textual and visual information. The models used for text were FastText, and for images, a pre-trained InceptionNet was used, combined through joint and common space fusion techniques. It achieved an accuracy of 86.92%. The paper did not consider applications apart from emotion classification and failed to take user-annotated data into account as well.

The authors of [8] analyzed social media data for disaster response using a multimodal deep learning model. They employed traditional models such as CNN to develop a combined representation from both text modalities and image modality of social media data. The proposed approach using multimodal data outperformed the existing unimodal approach, achieving an accuracy of 84.4%. However, the paper focuses on traditional models, neglecting more powerful newer models such as Vitb/16 or GPT-2.

In another study, [9] fused image and text data for UPMC Food-101 using BERT and CNNs models within a multimodal classification framework. BERT was utilized for extracting textual features, while InceptionV3 was employed for extracting visual features. A stacking technique was employed to perform the final multimodal classification using the UPMC Food-101 dataset. The models employed were BERT for text data and InceptionV3 for Image data, achieving an accuracy of 92.50%. However, the paper focuses on exploring the traditional CNN-based and transformer-based architectures, neglecting newer vision transformer and transformer-based language models.

Zou et al. [10] classified disaster images by integrating multimodal social media data through a fusion strategy that combines images and text to identify disaster-related images on social media. The experiment used VGG as an image feature extractor, FastText as a textual feature extractor, and a novel model to combine these extracted features for classification. The multimodal fusion was achieved using three fully connected layers and one SoftMax layer, resulting in an accuracy of 87.6%. However, the paper relies exclusively on a lightweight library; without considering the more powerful new age transformer models.

The research paper [11] introduces two late fusion mechanisms: weighting and meta combination methods for multimodal classification. Through these mechanisms, they have input visual features using CNN and textual features using Doc2 Vec and Bag of Words (BoG) to assess their impact on performance scores. For textual features, BoW was applied to the Random Forest (RF) algorithm, and Doc2 Vec was applied to the Support Vector Machine (SVM) algorithm. For images, the visual features extracted by AlexNet were fed into the SVM algorithm, combined through a stacking classifier, achieving an accuracy of 94.42%. However, the paper mainly focuses on the extraction and combination of features, neglecting deeper analysis or exploration of alternative model architectures.

Mouzannar et al. [7] identified damage in social media posts using multimodal deep learning, presenting a framework that combines pre-trained unimodal CNN architectures to independently extract textual and visual features, and a classifier is used for final prediction. For textual features, Kim's CNN architecture with Word2 Vec was utilized, and for images, InceptionV3 was employed. These were combined through FF meta-classifiers and DF classifiers, resulting in an accuracy of 92.62%. The proposed approach did not consider other ensemble models such as the RF classifier.

In their work, [12] detected hate speech in memes using multimodal deep learning approaches—a prize-winning solution to the hateful memes challenge. The focus is on multimodal hate speech detection within memes, using the VisualBERT model and then concatenating them using ensemble learning. The experiment results show high accuracy after implementing the proposed method. Both text and image models utilized VisualBERT, employing ensemble learning in combination, resulting in an accuracy of 76.5%.

Audebert et al. [13] used multimodal deep networks for text- and image-based document classification, proposing a neural network for multimodal classification that extracts text from images using OCR and FastText. Text was processed using a Multi-Layer Perceptron, while images were processed using MobileNetV2. These models were combined to form a similar representation space of dimension 128 in each of the baseline models.

Miller et al. [14] used multimodal classification using images and texts, proposing a methodology for integrating natural language understanding into image classification using associated metadata. The study implemented a multimodal image classification model that combines convolutional methods with natural language understanding of titles, tags, and descriptions through transfer learning. For text, Universal Sentence Encoder and BERT with DNN algorithm were used, while for images, InceptionV3, ResNet50 V2, and MobileNetV2 were employed and combined through a stacking classifier. The model achieved 46.60% and 70.56% Top1 and Top5 accuracy, respectively. However, the paper did not mention the information regarding the initialization of hyperparameters for each of the models used to extract features from texts and images.

Similarly, [15] conducted sentiment analysis of social media through multimodal feature fusion, proposing a model that cleans noisy text to extract textual features using an auto-encoder, and extracts visual features using the same auto-encoder with an attention mechanism. The fusion mechanism is based on learning through symmetry. For text, the model utilizes a Denoising Auto-Encoder, while for images, it employs an Attention-based Variational Auto-Encoder. These models are combined through a cross feature fusion module based on the attention mechanism, achieving an accuracy of 71.44%.

As can be seen in the aforementioned literature review, the exploration of classifying multimodal classification in the previous studies showed that the combinations of CNNs, BERT, and other models made it possible to improve prediction accuracy regarding audiences' reactions to the impact of different social media data types including emoticons, memes, and other emotional content such as videos, images, etc.

However, these studies often did not fully utilize the power of newer technologies such as Vision Transformers (ViT) and advanced language processing models such as GPT-2 that can extract potent linguistic abstracts and diverse visual elements because of their strength to capture semantic meanings and rich features. Our approach uniquely combines Vision Transformer (ViT) and GPT-2 models to exploit both textual and visual data, enhancing the balance and synergy between these modalities. Unlike traditional methods that predominantly focus on either modality or utilize less advanced models [5, 8], our ensemble model introduces a more sophisticated integration strategy. This not only leads to significant improvements in accuracy—surpassing the baseline Convolutional Neural Network (CNN) models by 13.23% in predictive accuracy—but also ensures a robust classification system that can adapt to the dynamic and varied nature of disaster-related data. The novelty of our transformer-based approach lies in its capacity to integrate heterogeneous data modalities seamlessly into the decision-making process, offering a critical advancement over existing approaches like those discussed in [9] and [10]. This methodological innovation represents a significant step forward in multimodal data analysis, pointing the way for future research to further harness the combined strengths of advanced transformer models.

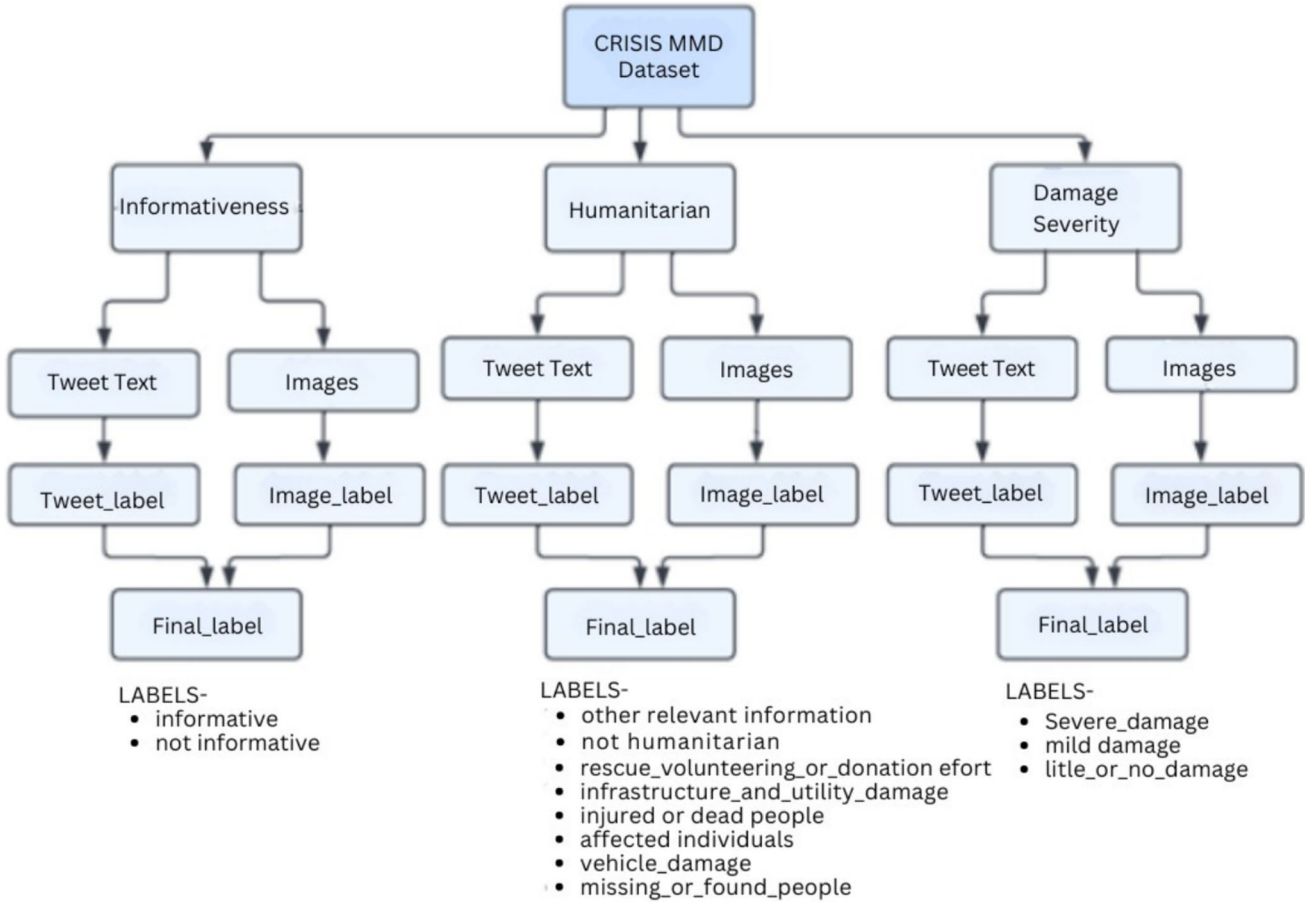

Fig. 1 Categorization of crisis MMD dataset according to different tasks

3 Data and Methodology

3.1 Data Acquisition

Social media has been the most popular medium of real-time data collection. Since many users tend to post crucial information on social networking sites, Twitter being the most popular one, we adopted the recent multimodal dataset curated by Alam et al. for their study on "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters" [16]. The Crisis MMD dataset is a multimodal dataset collected from Twitter containing tweets and related images of natural disasters in 2017. The dataset contains the data from seven natural disasters in different regions of the world, including all types of calamities such as earthquakes, floods, wildfires, and hurricanes. The data have been categorized into three types of annotations for three different tasks, as shown in Fig. 1, including:

1. Informative and not informative
2. Humanitarian categories
3. Damage severity assessment

The classification task performed in this study belongs to the first category predicting whether a particular tweet containing image and text is informative or not. This task checks if a specific tweet text or image acquired while a calamity has taken place may be used for relief help. It is regarded as an "informative" tweet if both image and text associated with it are valuable for relief help; otherwise, it is deemed as a "not-informative" tweet [16]. The dataset was collected from Twitter using the Twitter API with the help of trending hashtags and specific keywords during disaster events. Figure 2 lists the keywords used and the data collection period for each event [17].

The dataset consists of 16,058 tweets and 18,082 images related to natural disasters. The distribution of data used for training can be understood in Fig. 3a, b. Figure 3a indicates the distribution of informative and not informative labels for textual data, whereas Fig. 3b illustrates a similar distribution for image data. Specifically, the textual data consisted of 9638 informative and 3970 non-informative labels, whereas the visual data included 7059 informative and 6549 non-informative labels.

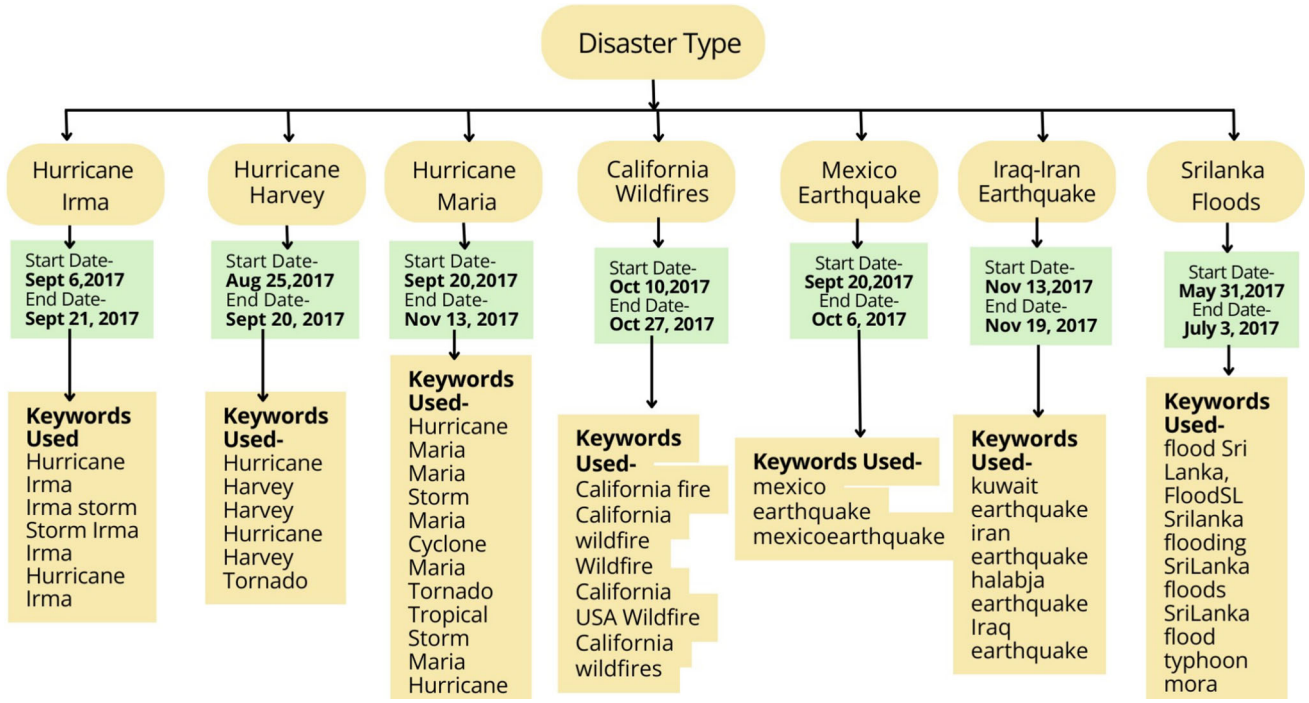

Fig. 2 Categorization of CrisisMMD dataset according to event names, keywords used for data collection, and data collection period

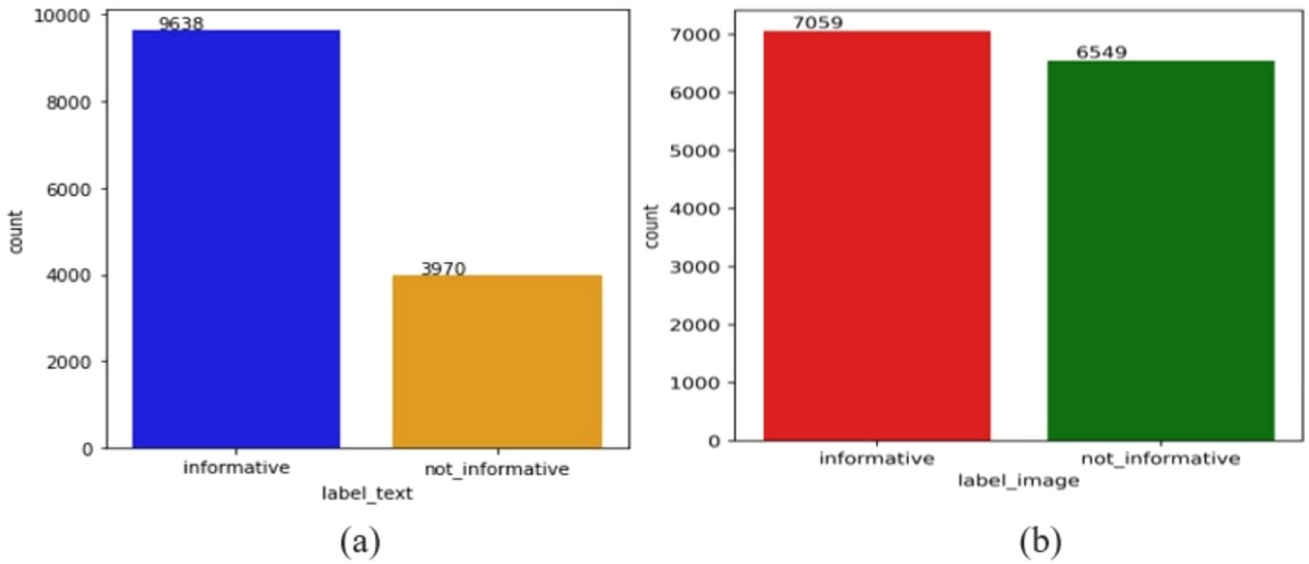

Fig. 3 Bar chart distribution of: **a** text target labels and **b** image target labels

Since this paper focuses on classifying the informativeness of the text and image data, the Crisis MMD dataset was used for the first task. The informative folder consisted of 13,608 records for the training set with nine non-null columns that included event name, tweet id, image id, tweet text, image, label, label text, label text label image, and label text image. The text data was trained on the tweet text column using the target label as the "label text" column, and the image data was trained using the column "image" consisting of image paths with "label image" as the target column. The validation and test data consisted of the same number of columns with 2237 records, as shown in Table 1.

Table 1 Label distribution of task A (informativeness)

Informativeness (Task A)			
Label	Tweets (Text)	Image	Final label
Train dataset			
Informative	9638	7059	8341
Not informative	3970	6549	5267
Total records	13,608	13,608	13,608
Validation dataset			
Informative	1612	1164	1407
Not informative	625	1073	830
Total records	2237	2237	2237
Test dataset			
Informative	1612	1151	1373
Not informative	625	1086	864
Total records	2237	2237	2237
Final count	18,082	18,082	18,082

3.2 Data Preprocessing

The initial stages of multimodal data classification involve taking suitable features from different modalities and merging them to improve the prediction performance. Various techniques for preprocessing and extracting features from texts and images are introduced by data scientists based on their extensive research work with multimodal data. These techniques include using models such as VGG-19, ResNet50, InceptionV2, Xception, and DenseNet [4], as well employing CNNs for text modality, VGG-16 architecture for extracting high-level features from images [8], and the FastText framework [10], among others. In our implementation, we decided to apply some of these preprocessing and feature extraction techniques. Below, we discuss a few of the techniques that we have focused on.

3.2.1 Text Preprocessing

The textual data in the CrisisMMD dataset often contains noise in various forms, such as punctuation, stop-words, unique symbols, and numerical values, which can impede prediction accuracy. To address this, preprocessing of the textual data began with converting tweets to lowercase and removing noisy elements, including hashtags, URLs, HTML references, placeholders, non-letter characters, and punctuation, as depicted in Fig. 4. Additionally, predictive models cannot directly comprehend natural language text; they require numerical representations. Therefore, the GPT-2 transformer model automatically tokenized the text data using byte pair encoding followed by label encoding on the text dataset giving us preprocessed text data.

- *Byte pair encoding* The byte pair encoding is a method of sub-word tokenization relying on word- and character-based tokenization [18]. The breaking of raw text into small pieces of text is known as tokenization, which helps models understand the context of the natural language text. The sub-word-based tokenization splits the rare words into smaller meaningful sub-words; for example, the term "boy" is split into "boy" and "s." Byte pair encoding is the simplest data compression method because it replaces the frequent byte pair with a new byte. As a result, it compresses the data [19].

3.2.2 Image Preprocessing

Preprocessing of images aims to improve image quality by suppressing undesired distortions and enhancing specific features relevant to the particular application at hand. These features may vary according to the application being addressed. In the proposed experiment, we aim to classify the images as informative or non-informative. Thus, certain preprocessing steps are necessary before providing the images to the model for prediction. The preprocessing techniques applied to the CrisisMMD image dataset include resizing, vertical flipping, horizontal flipping, cropping, conversion of the image to a tensor, and normalization, as illustrated in Fig. 5.

- *Resizing* Resizing images was necessary for the efficient performance of Vit Base 16 because resizing images affects the model's training time and performance significantly [20]. Since the CrisisMMD image dataset contained images of various sizes, resizing them to a uniform size was necessary to ensure consistency in the number of extracted features across all images [20]. The images in the CrisisMMD dataset were resized to a width size of 224 and height value of 224. An example of resizing is shown in Fig. 6 on one of the images from the CrisisMMD dataset.
- *Vertical and horizontal flipping* Applying flip augmentation gives more information for the model to learn during the training process [21]. It is a technique in which an image is rotated on a horizontal or vertical axis to reduce overfitting and introduce diversity in the data [22]. The CrisisMMD images were flipped horizontally and vertically with a probability of 0.3, as shown in Fig. 7a, b.
- *Cropping* Image cropping is a manipulation process to improve the overall composition of the image by eliminating those parts of the image which are not desirable [23]. In this process, an image is taken, followed by creating a random subset of the original image. For the ViT Base 16 model to accurately predict the images as informative or non-informative, the images in the CrisisMMD dataset were cropped to size 224, as shown in Fig. 8. Cropping

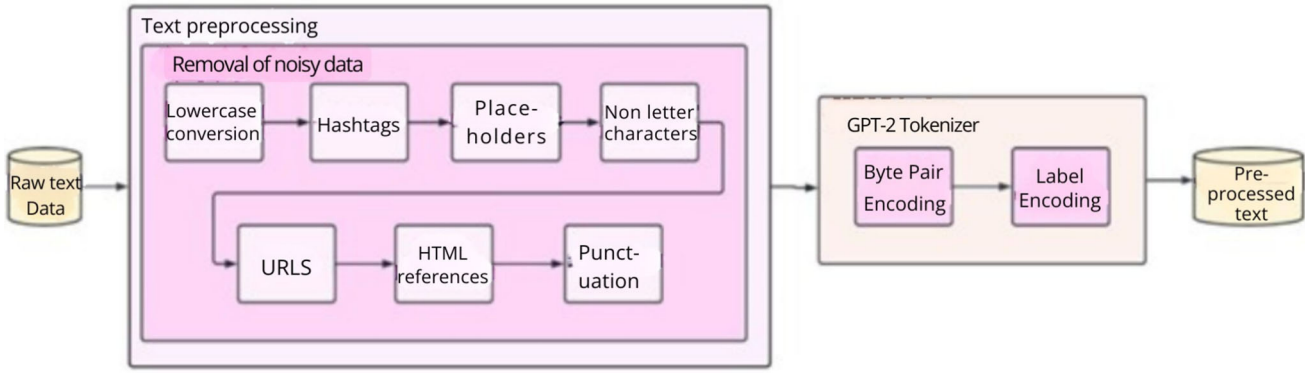

Fig. 4 Preprocessing pipeline of CrisisMMD text data

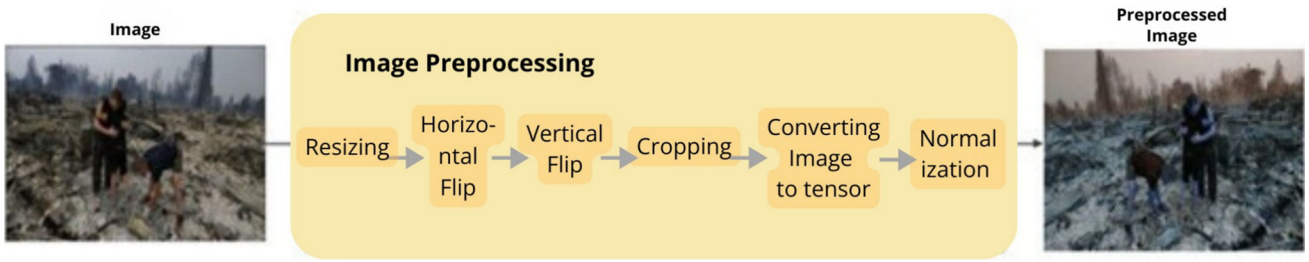

Fig. 5 Data preprocessing pipeline of CrisisMMD images

was performed to adjust the outside edges of the image to improve framing or composition by changing the size or aspect ratio and drawing the model's attention to the image subject.

- *Normalization* Normalization is done to change the pixel intensity values of the images. In general, normalization transfers numerical features into a standard range of values. Pixel values in an image range from 0 to 256, where each number represents a color code. When an image is normalized, it divides the high numeric values by 255 into numeric values ranging from 0 to 1, resulting in equal distribution of data. In this experiment, the CrisisMMD image dataset was converted to tensor values before it was given for normalization, and after normalization, the images appear similar to Fig. 9.

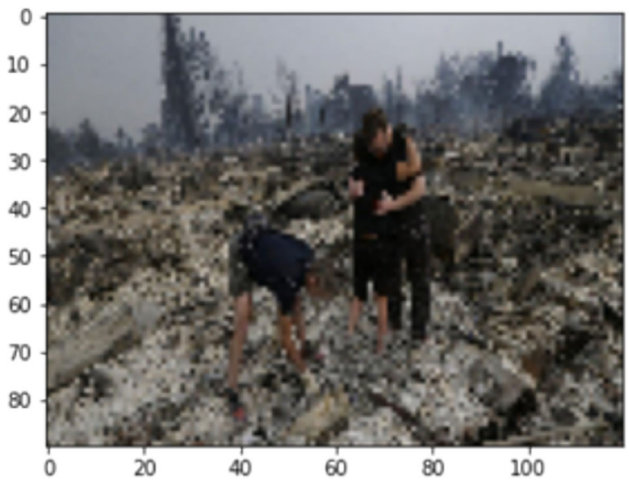

Fig. 6 Image after resizing

4 Models

4.1 Transformer Models

Transformer models [24] have lately exhibited outstanding performance in many areas, including text categorization, machine translation [25], image classification, and question answering. As opposed to recurrent networks, such as long short-term memory, transformers enable the modeling of lengthy dependencies. It is also capable of processing sequences in a parallel manner. Unlike convolutional networks, they are designed with minimal inductive biases. They have great scalability for large-sized datasets. These advantages have made Transformer models the talk of the town [26].

The architecture of transformer models is based on the mechanism of self-attention, which studies the relationships between parts of a sequence. It enables the recording of "long-term" data and dependencies between sequence pieces. The transformer has an encoder–decoder structure. The encoder's job is to convert an input sequence to a

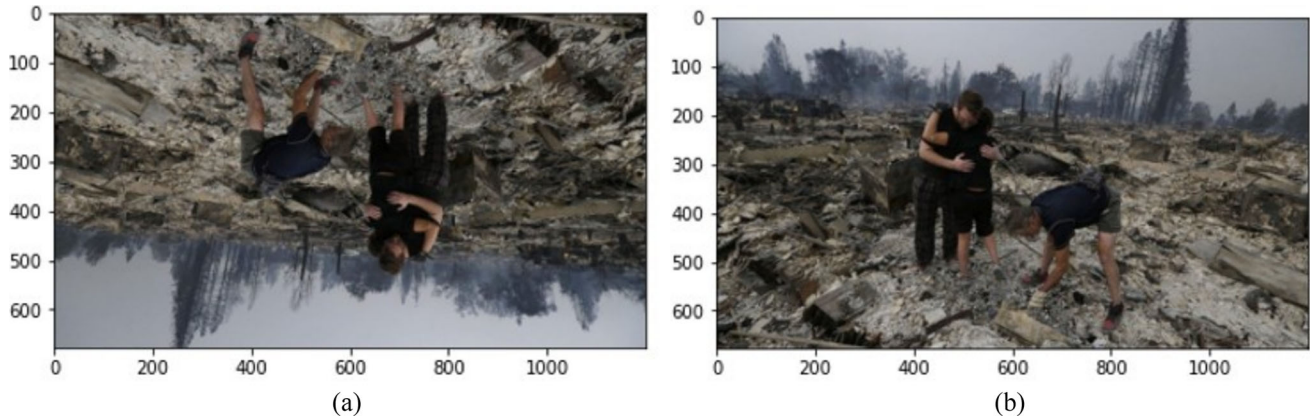

Fig. 7 Image after applying **a** vertical flip and **b** horizontal flip

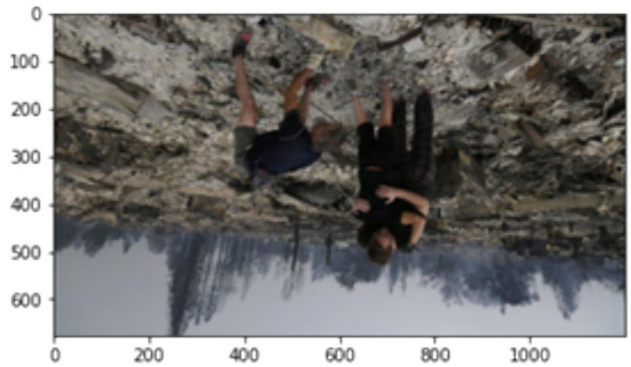

Fig. 8 Image after cropping

Fig. 9 Image after normalization

sequence of continuous representations, which is then fed into a decoder. The encoder output is then combined with the decoder output to form an output sequence.

Feed-forward and recurrent networks have used attention models earlier [27, 28]; however, transformers stand out from the rest because they can process sequences parallelly for optimization. Pre-trained transformers on massive datasets are very common since it allows them to learn the problem's internal structure, unlike the convolutional neural network models [24, 29–31]. Fine-tuning of the learned representations is performed in a supervised way on downstream tasks to get favorable results. These reasons led us to employ a vision transformer model and a transformer-based language model-GPT-2 for multimodal classification in our proposed approach.

4.1.1 GPT-2: Text Modality

The GPT-2 model was created by Open-AI, trained on text data from 8 million websites with 1.5 billion parameters (ten times more than what the original GPT had). It is a decoder transformer that uses the last token of the input sequence to anticipate the next token that should come after it. This signifies that the last token of the input sequence provides all the required data to make the prediction. The GPT design leverages attention instead of earlier recurrence- and convolution-based architectures to create a deep neural network. Our proposed approach uses the GPT-2 model for tweet classification. Given that GPT-2 represents one of the latest advancements in transformer-based language models, we sought to explore its synergy with another contemporary image classification model known as ViT Base 16. In contrast with BERT, a baseline transformer model, decoder blocks are used to construct GPT-2 transformers providing only one token at a time as an output. Each created token is added to form a sequence of inputs which becomes the input for the model in the further step.

The working of the GPT-2 models, as shown in Fig. 10, can be understood as follows:

1. *Input embedding* The initial step in the GPT model is to embed the input followed by positional encoding to ensure the ordering of the words. The words are then passed in a sequence to the transformer blocks.
2. *Decoder* The first block processes the token through the self-attention block. The self-attention block identifies

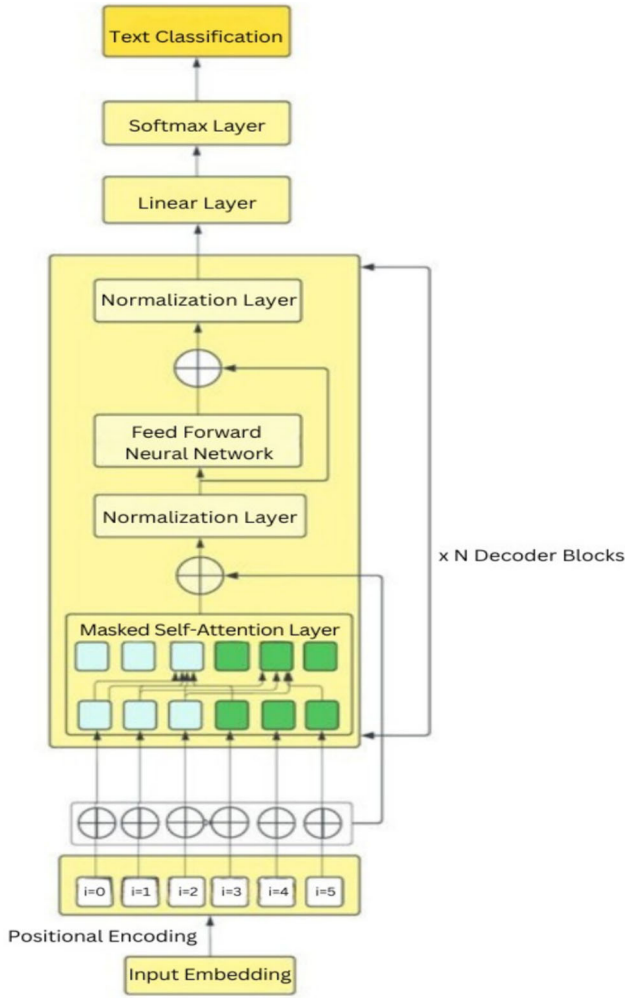

Fig. 10 GPT-2 architecture

the critical word from the text that needs to be focused on. The exact process continues with many other decoder blocks. Lastly, the result of all these decoder blocks is concatenated to provide the final output, which is then fed to the final feed-forward layer. Besides, there is a normalization layer between each self-attention and feed-forward network to improve the learning process of the model.

3. *Feed-forward neural network* The last step is to convert the output to a probability distribution. Each number represents the likelihood of the given token being the correct one using the SoftMax layer. This process continues for the entire data, after which a random sampling method is employed to pick the correct token. That token is converted to a word from the dictionary created, and the final label of the text is predicted as the output

4.1.2 ViT Base 16: Image Modality

In recent years, CNN has become the most popular model for visual identification and classification [32–37]. On the other hand, recent research suggests that successful recognition and classification models can be built without convolutions [38]. Vision Transformer (Vit) is the most representative effort in this direction. Unlike traditional convolutional techniques, which process pictures pixel by pixel, the Vision Transformers (Vit) handle images as a sequence of patch tokens. It uses multi-head self-attention to recombine and process these patch tokens at each layer depending on the relationships between each pair of tokens, as shown in Fig. 11. In this manner, Vit models can generate a general representation of the overall image. As a result, Vision transformer models have been successfully used in many applications, including image recognition [29, 39], object detection [40, 41], image classification [32], image generation, and visual question answering [34, 35], among others.

Its architecture is transformer-based, wherein it splits the images into patches of fixed size followed by embedding and positional encoding, which are then received by the transformer encoder as input, as shown in Fig. 11. The self-attention layer allows us to embed information globally throughout the entire image. It is a computational primitive that helps a network learn the scaling and symmetry contained in incoming data by quantifying paired entity interactions. The model utilizes the data used for training to encode the respective placement of image patches which helps in recreating the structure of the image.

Additionally, leftover connections are provided after each block because they promote a hassle-free flow of items through the network. The classification head in image classification is implemented by the Multi-Layer Perceptron (MLP) layer, a collection of linear transformation layers.

The entire architecture of the Vit Base 16 models, as shown in Fig. 11, is explained below in a stepwise manner:

- Creates patches of fixed size from the input image.
- The patches of the image are then made flat.
- Lower-dimensional linear embeddings are created from these flattened image patches.
- Positional embeddings are included.
- The sequence is fed into a state-of-the-art transformer encoder as an input.
- Layer Norm (LN) is added because there are no new dependencies between the training images. As a result, the training time and overall performance are improved. Each block has an LN in front of it.
- Image labels are pre-trained before putting them through their paces on a large dataset.
- Image classification is then fine-tuned on the downstream dataset.

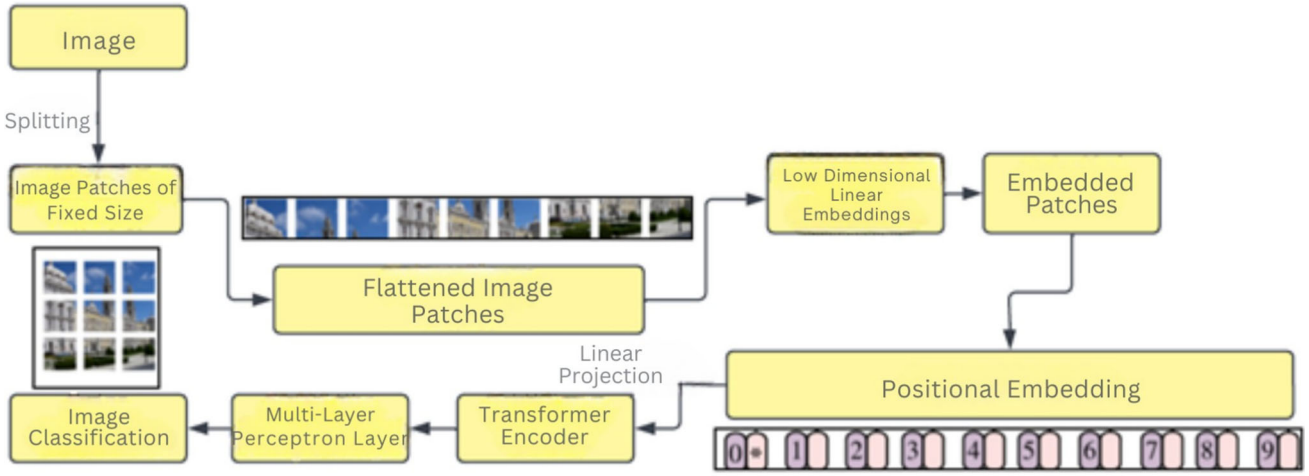

Fig. 11 Vision transformer architecture

This paper focuses on VIT-B/16, the "base" variant with a 16×16 input patch size. Vit-16 demonstrates exceptional performance on being trained on ample amounts of data, surpassing the performance of CNN models, which have fewer computational resources, leading us to employ this particular transformer model for the proposed image classification task.

4.2 Ensemble Model

An ensemble model is a machine learning approach that involves combining multiple other models in the prediction process, with these models referred to as base estimators. It serves as a solution to overcome technical challenges such as high variance, low accuracy, and feature bias encountered when building a single estimator. Random forest is one such ensemble model in which the ensemble approach is bagging, and the individual model is a decision tree.

4.2.1 Random Forest (RF) Classifier

RF is a modern ensemble learning technique for classification. It models discrete data using classification trees [42] and classifies the data by generating several Classifiers to improve prediction accuracy [43]. If a single classifier is used to classify massive amounts of data, it may lower accuracy. Because a considerable amount of data must be categorized [44], we picked RF as the ensemble model to make final predictions in our proposed approach.

5 Proposed Architecture

The study presented in this paper follows the architecture shown in Fig. 12. The idea is to develop a classification system for multimodal data that can classify the images and text related to natural disasters based on informativeness to help fasten the disaster response. We have used the latest architectures in text and image classification to achieve the goal, i.e., Vit Base 16 and GPT-2 transformer models. The proposed architecture adopts a novel pipeline with the help of these two transformer models. Both the models are first trained separately on the CrisisMMD dataset, followed by the concatenation of the prediction outputs of both models using the ensemble model-Random Forest Classifier to yield the final prediction for the provided input.

The multimodal classification was performed using the CrisisMMD dataset containing records related to the informativeness of the tweets. GPT-2 model and the Vit Base 16 model have been trained on the text and image data. The data were first preprocessed before feeding the input to the respective models using suitable techniques. The twitter text is processed and transformed into clean data by removing the noise, including hashtags, URLs, HTML references, non-letter characters, punctuations, placeholders, and Twitter mentions. The entire text is converted into lowercase before passing the data to the tokenizer. The GPT-2 tokenizer then tokenizes the text and performs label encoding and padding on the data. Similarly, image preprocessing is done by resizing them into the proper format and flipping them vertically and horizontally. The images are then cropped and converted to tensor form, followed by normalization. The GPT-2 model performs embedding (text and positional embedding) on the input text as the first step, followed by the self-attention layer that assigns weight to each token and identifies the association between each data point. The process continues for all

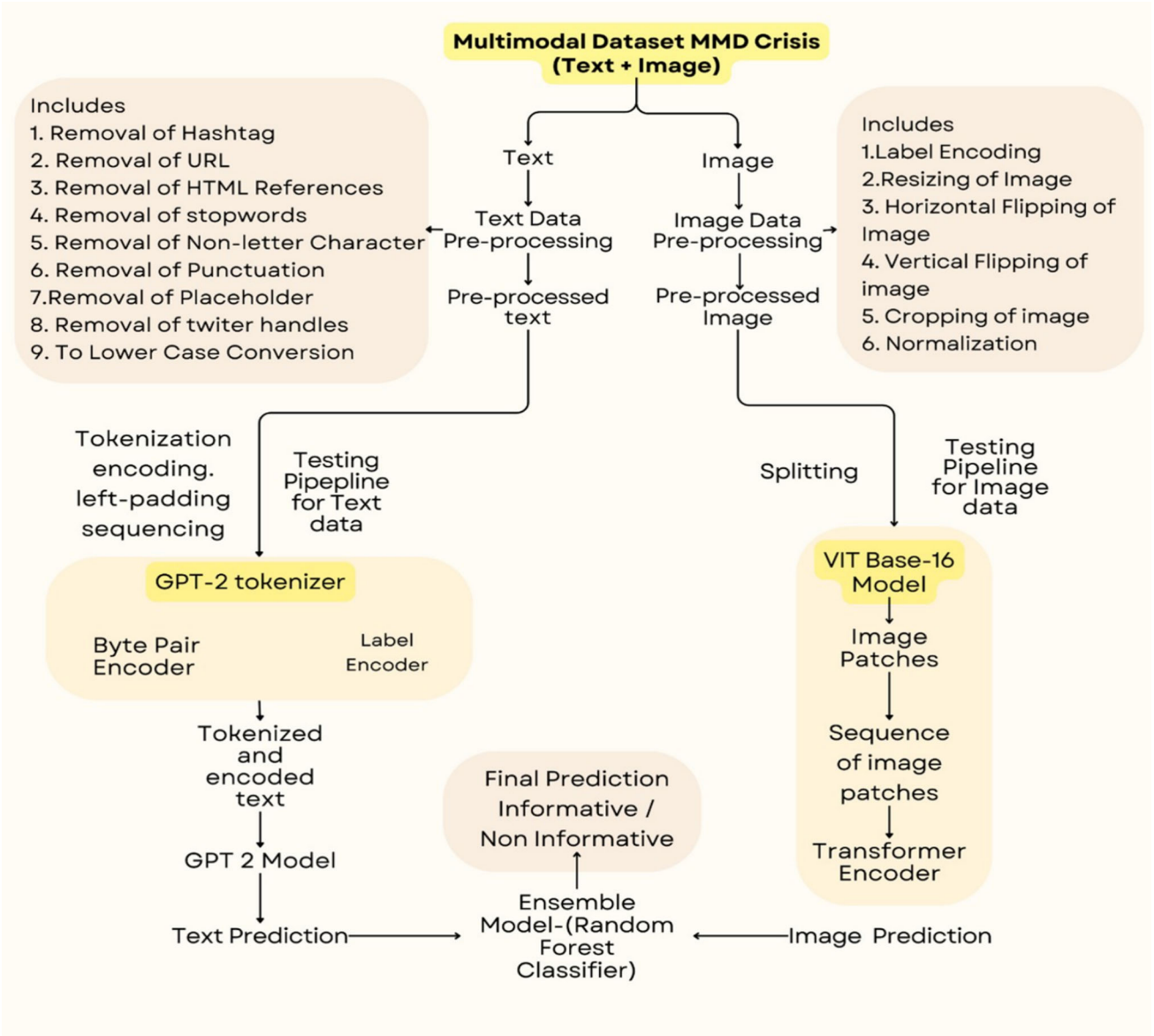

Fig. 12 Proposed system architecture

the decoder blocks present in the model, passing the output token to the feed-forward neural network. It gets converted into a probability distribution over there. The final predicted label of the input text is produced as an output. Similarly, the Vit Base 16 model takes the preprocessed image and splits it into patches which are treated as inputs to the transformer blocks. The patches are fed in sequence as a list of vectors having dimensions 16 by 16 to the transformer. The transformer model passes these vector images through an encoder block that comprises the Multi-Head Self-Attention (MSA) and MLP layer. The MSA layer attaches attention to each patch and provides the concatenated output. This output is an input to the MLP layer that provides the output label of the input image. The predictions from both (GPT-2 and Vit Base 16) models are concatenated with the help of an ensemble random forest classifier model. The classifier combines predictions from both models to predict the provided data's final class label (either as informative or not informative).

6 Experimental Setup

6.1 Parameter Settings

This section covers the implementation details of our multimodal classification approach by defining the hyperparameters used while training the models. The image dataset was

Table 2 Vision transformer hyperparameters

Parameter name	Data type	Default value	Chosen value
Epochs	Integer	1	10
Learning rate	Float	le−06	2e−05
Batch	Integer	16	16
Image size	Integer	224	224
Seed size	Integer	1	1001
Optimizer	None	Adam	Adam

Table 3 GPT-2 hyperparameters

Parameter name	Data type	Default value	Chosen value
Epochs	Integer	1	2
Learning rate	Float	5e−05	2e−5
Batch	Integer	16	128
epsiLon	Float	2e−05	2e−8
Seed size	Integer	1	130
Optimizer	None	Adam	AdamW

trained on Tensor Processing Unit (TPU). Similarly, Graphical Process Unit (GPU) was the hardware resource we chose to train the text model. The fusion of predictions by both the transformer models was loaded on the CPU and concatenated with the help of an ensemble model.

6.1.1 VIT (Vision Transformer) Base 16 Model

We opted for the base version of the vision transformer with an image patch size of 16 * 16 for the imaging modality due to its better performance in image classification tasks. The environment was set up by installing and importing torch xla to connect the PyTorch deep learning framework to TPU, with the entire training performed on 8 TPU cores. The image classification was performed using PyTorch version 1.7 and Timm library. The Vit Base 16 was pre-trained on the ImageNet dataset [39]. We leveraged this pre-trained Vit Base 16 model's existing weights and trained our model on images resized at 224 * 224, cropped at 224, horizontally and vertically flipped at a probability of 0.3, and normalized using the mean and standard deviation of default values.

These images were then converted into tensors for feeding them to the model. It was optimized using an Adam optimizer with a momentum of 10 epochs, a learning rate of 2e−05, and a batch size of 16, as shown in Table 2. The seed size was set to 1001 for better reproducibility of the results. The cross-entropy loss between the input and targets was calculated using a cross-entropy criterion. We went for a smaller patch size, i.e., 16 * 16, to improve the quality of the resulting features [45]. Additionally, the epsilon value was set to 1e−06 to ensure that the model does not encounter a divide by zero error.

6.1.2 GPT-2 (Generative Pre-trained Transformer) Model

For text modality, the GPT-2 model was used for training. The environment was set up by installing and loading the transformer models from Github [46], consisting of three deep learning libraries, Jax, PyTorch, and Tensorflow, followed by installing the helper functions to avail the Python libraries useful in NLP tasks. The training was done on the Graphical Processing Unit (GPU) provided by Google collab with a seed size of 130 to optimize the training result's reproducibility. For better training, the batch size was set to 128, and padding of the text was allowed to 40. The preprocessed text fed to the model was tokenized by the GPT-2 tokenizer that also performed label encoding of the texts. The model was trained on two epochs using AdamW as the optimizer with 2e−5 as the learning rate for all three data loaders (train, text, and valid data loader) having an epsilon value of 2e−8 as shown in Table 3.

Table 4 Random forest classifier hyperparameters

Parameter name	Data type	Default value	Chosen value
n estimators	Integer	10	100
max leaf nodes	Integer	None	3
max depth	Integer	None	2
min_samples_leaf	Integer	1	2
criterion	String	"Gini"	"Entropy"

6.1.3 Random Forest Classifier Ensemble Model

The Random Classifier Ensemble model was then used to combine the image and text predictions made by transformer models to make a final prediction. For the random forest classifier model, the "label image" and "label text" columns were input features, and the "label" column was the output feature for training purposes. In this experiment, the value of n estimators was set to 100. Hyperparameter values of max leaf nodes, max depth, and min samples leaf were set to 3, 2, and 2, respectively, as shown in Table 4. The criterion for measuring the quality of a split was set to "entropy." We did not tune the hyperparameters using any specified technique. As a result, this aspect can be analyzed further in the future research. The reason for choosing a random forest classifier as the final ensemble model for our approach was to achieve better accuracy through cross-validation.

The ensemble method was chosen for its ability to combine the strengths of multiple models and to improve the

result in terms of accuracy and robustness. This approach was prioritized to address the high stakes of decision-making in scenarios such as disaster response. Future research could explore integrating explainability techniques, such as SHAP or LIME, into the ensemble model.

6.2 Evaluation Metrics

Evaluating a model is essential for building effective transformer models and achieving considerable accuracy during model training. Building these models requires constructive feedback from evaluation metrics to improve the transformer models' performance and continue until a desirable result is achieved. The four evaluation methods used in this experiment are accuracy, confusion matrix, classification report, and ROC-AUC curve.

- *Confusion matrix* The confusion matrix is a performance measurement for classification tasks that are represented in a table with four different values—true positive (TP), true negative (TN), false positive (FP), and false-negative (FN), as shown in Fig. 13, wherein the target classes in our multimodal classification are informative and non-informative. The confusion matrix is functional for measuring recall, precision, accuracy, and the ROC-AUC score.
- *Accuracy* The accuracy metric, which determines the fraction of correct predictions to the total number of samples as defined in Eq. (1) [47]. It answers the question of what percent of the model's predictions were correctly predicted where it sees the number of true positives and true negatives compared to the entire sample as shown in Eq. (1).

$$\text{Accuracy}(a) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

- *Classification report* Using the values from the confusion matrix, the precision, recall, F1-score, and accuracy are all calculated and displayed on the classification report. Precision and recall measure the true positives in the samples, but precision explores the number of TP to the aggregate of TP and FP, as shown in Eq. (2). In contrast, recall looks at the count of TP to the total of TP and FP, as mentioned in Eq. (3) [47]. The F1-score is the weighted harmonic mean of precision and recall seen in Eq. (4).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

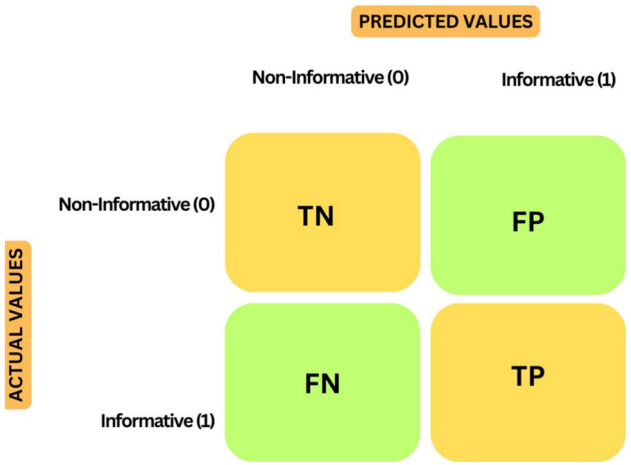

Fig. 13 Confusion matrix table

- *AUC-ROC curve* The AUC-ROC curve was used to check the performance of the proposed binary classification problems. AUC stands for the area under the curve, while ROC stands for receiver operating characteristic. The curve plots two parameters, which are the true-positive rate (TPR) and false-positive rate (FPR) [48]. The TPR is the ratio of true-positive values to the positive values as per Eq. (5), while the FPR is the ratio of false-positive to the negative values, as shown in Eq. (6).

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{6}$$

7 Results and Discussion

This section discusses the results of implementing the transformer and ensemble models in detail for both image and text modalities. In summary, we analyzed the multimodal CRISIS MMD dataset for predicting the informativeness of the crisis-related data containing text and images. For image modality, we went for Vit Base 16 to classify and predict whether the images were informative or not. On the other hand, we chose GPT-2 as the classification model for text modality. The predictions were then combined to get a final prediction using the Random Classifier Ensemble Model.

Table 5 presents the performance results achieved for all three models. For Image Classification, we achieved an accuracy of 81.9% using the Vit Base 16 Transformer Model, whereas, for text classification, the GPT-2 Model gave an accuracy of 80.7%. For final predictions, we opted for a random classifier model to concatenate these predictions in

Table 5 Dataset evaluation concerning transformer models and convolutional neural network models

Task	Model	Accuracy (%)
Image classification	Vit-16	81.90
	ResNet-50	67.50
Text classification	GPT-2	80.70
	CNN	72.06
Final prediction	Random forest	84.66
	CNN + ResNet-50	71.43

which a performance score of 84.6% was achieved. In all the three cases, the implemented models outperformed the CNN Models by 14.4%, 8.64%, and 13.23%, respectively. Hence, the overall performance of the informative classification transformer models was better than the existing neural network models due to the simple nature of the transformers and their ability to model long dependencies and support parallel processing [26].

Since our study focuses on multimodal data classification, we visualized the performance of different models with the existing state-of-the-art ones. Figure 14 shows the accuracy of the informative classification task concerning the image, text, and multimodal models. Comparative performance analysis with the already existing CNN models is shown in Fig. 14.

In our study, we tested the method on datasets ranging from 1000 to 100,000 samples, and it showed consistent performance improvements over baseline methods. Scalability testing is important to validate the proposed approach and to check the robustness of the findings on larger datasets in the future work. Access to larger datasets and computational resources is a challenge but online GPU workstations can provide the necessary infrastructure and enhance the scalability.

Transformer models require a lot of data [49]. The dataset we used for this study comprises around 39,690 records. Finding concatenation procedures that better capture crucial details from text and image data are complex. Hence, the random forest classifier ensemble model was designed in this direction to concatenate the predictions of the image classification model and the text classification model into a single shared representation. We further analyzed the confusion matrices of the three models to see how they performed, as shown in Fig. 15.

Our text-only models missed 227 cases, whereas, for the image-only models, the false-negative value was 110, according to Fig. 15a, b. Figure 15c shows that the multimodal model missed none. When the computer says "not informative," but the actual label says "informative," this is what happens. The image classification model outperformed the text classification model, but when the text and image modalities were mixed in the multimodal instance, the values fell dramatically (i.e., from 110 to 0). Another primary consideration is when a message is predicted as informative but is not informative. The values for false-positive scenarios were 177 for image-only, 320 for text-only, and 360 for multimodal. We see no benefits from the multimodal strategy, as we saw in the false-negative scenario.

AUC-ROC statistics are well-known for measuring a model's ability to discriminate between classes [50]. As a result, we chose to present the experimental results in terms of the AUC-ROC curve, the most widely used evaluation metric, elucidating why such a pipeline was chosen. Since the area under the ROC curve represents the model's accuracy, Fig. 16a clearly shows that the VIT Base 16 model for image classification was quite accurate. The area under the curve is approximately 0.8.

The image-only model gave a ROC-AUC score of 0.81, which is another factor to indicate that the model was good at predicting informative and not informative labels accurately. The text classification GPT-2 Model was pretty accurate, as shown in Fig. 16b, with a ROC-AUC score of 0.70. The final text + image classifier model was the most accurate, with the area under the curve close to 0.85, as shown in Fig. 16c. The ensemble model fused all of the predictions at the decision level, resulting in the best ROC-AUC score of 0.87. As a result, the ROC curves and AUC values increased considerably when an ensemble learning model was employed to merge image and text predictions.

Figure 17 shows the classification reports for all three models: image, text, and image + text. In this study, we have evaluated them using this particular evaluation metric since the accuracy is not enough to evaluate the performance of these models. Even though accuracy is mainly used to judge model performance, it can suffer from anomalies when classes are imbalanced. The Vit Base 16 model's corresponding measures have better scores than the GPT-2 model. As shown in Fig. 17a and Fig. 17b, the precision scores of the Vit Base 16 model are 0.07% greater for non-informative and 0.02% greater for informative compared to the precision scores of the GPT-2 model. In addition, the recall scores of the Vit Base 16 model are 0.36% greater for non-informative and 0.13% less for informative compared to the GPT-2 model. However, the random forest classifier model gave a better recall value than the Vit Base 16 model and GPT-2 model, as shown in Fig. 17c.

Meanwhile, the F1-score of the Vit Base 16 model is 0.24% greater for non-informative and 0.05% less for informative contrasted to the F1-scores of the GPT-2 model. Figure 17c shows that the precision, recall, and F1-scores for informative classes in the random forest classifier model are 0.72, 1.00, and 0.83, respectively. The final model has

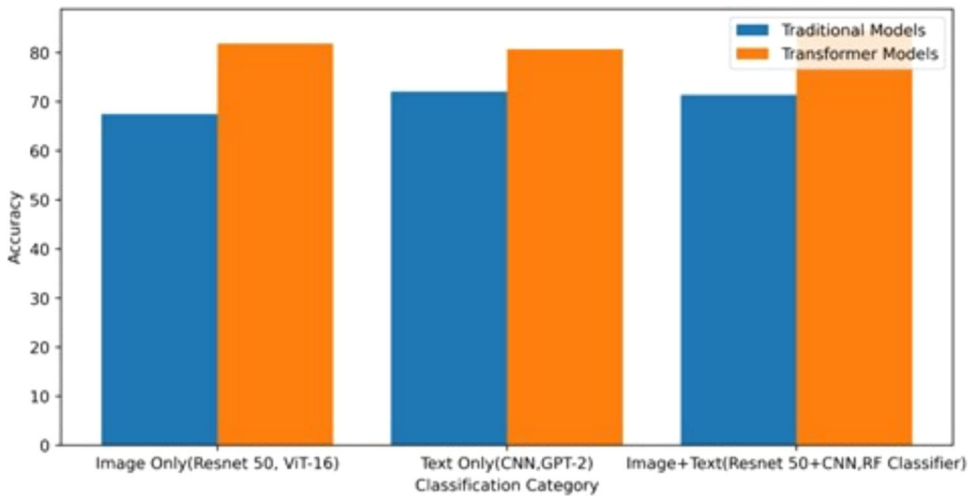

Fig. 14 Comparison of transformer models with traditional models in terms of accuracy

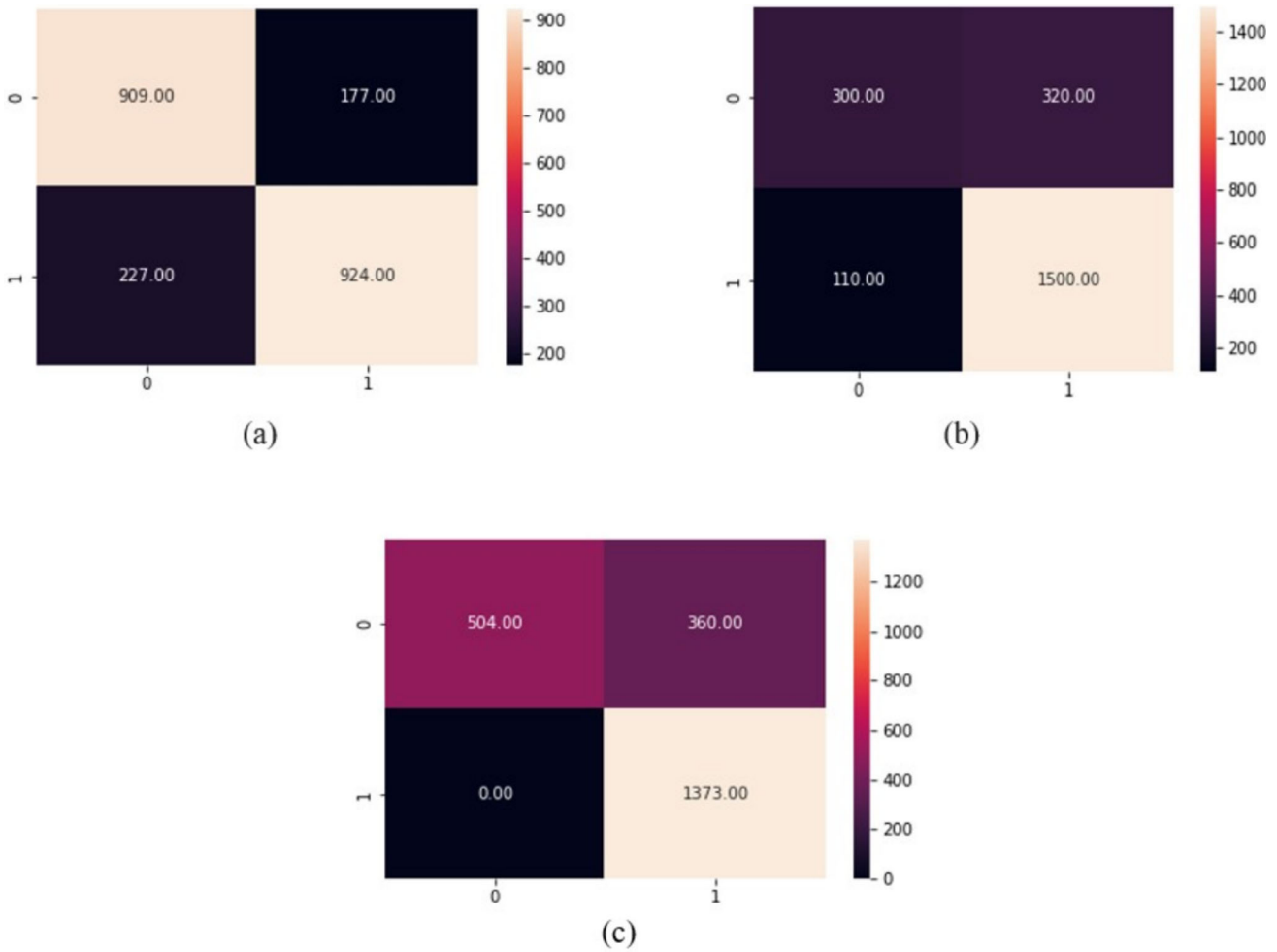

Fig. 15 Confusion matrices plotted for **a** Image-only model [VIT Base 16], **b** Text-only model (GPT-2), and **c** Image + text model (Random forest classifier model)

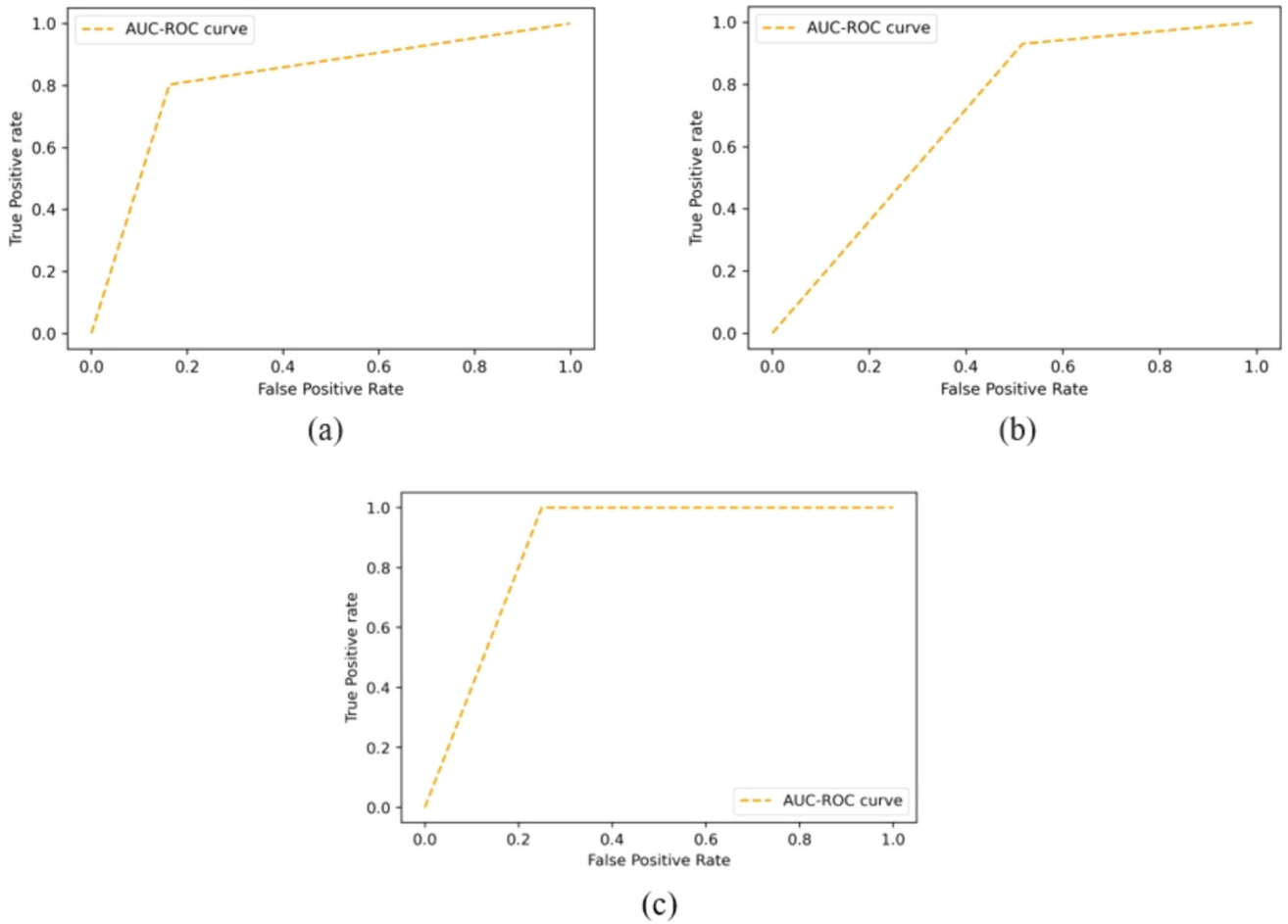

Fig. 16 ROC-AUC curve concerning **a** Image-only model [VIT Base 16], **b** Text-only model (GPT-2), and **c** Image + Text model (Random forest classifier model)

the best F1-score among all the three, indicating that it successfully integrated data from both modalities to arrive at an accurate prediction. These details, we feel, add to our understanding of the suggested multimodal approach for the classification of crisis-related multimodal tweets.

For this study, the focus was on achieving high accuracy and demonstrating the potential of combining these models, rather than optimizing for deployment. To address these challenges, future work could explore techniques such as model compression, quantization, and knowledge distillation to reduce computational overhead. Additionally, alternatives like lightweight transformer variants (e.g., MobileViT or DistilGPT) could be investigated to make the models more suitable for deployment on edge devices. The theoretical analysis of the time and space complexity of the proposed method demonstrates that the algorithm is efficient for large size datasets under typical conditions but may require further optimization to handle worst-case scenarios effectively. Since our model is a transformer-based architecture, it requires more computational power. The time complexity of O (N 2 d) requires O (N 2) space.

8 Conclusion and Future Work

Multimodal classification has garnered significant attention among researchers, with numerous experiments conducted using diverse datasets to achieve notable results in classification models. However, the approach used to attain the task has been recurrent in many previous works. Hence, in this paper, we aim to find a solution for multimodal classification with the help of transformer models such as vision transformer and GPT-2 to achieve improved results as compared to state-of-the-art models. The multimodal model achieved 84.66% accuracy outperforming the traditional CNN models by 13.23%. The model achieved better results than the unimodal ones.

The achieved results can be improved with further research that we aim to examine in the future. This study can

	precision	recall	f1-score	support
0	0.80	0.84	0.82	1086
1	0.84	0.80	0.82	1151
accuracy			0.82	2237
macro avg	0.82	0.82	0.82	2237
weighted avg	0.82	0.82	0.82	2237

(a)

	precision	recall	f1-score	support
0	0.73	0.48	0.58	625
1	0.82	0.93	0.87	1612
accuracy			0.81	2237
macro avg	0.78	0.71	0.73	2237
weighted avg	0.80	0.81	0.79	2237

(b)

	precision	recall	f1-score	support
0	1.00	0.75	0.86	1373
1	0.72	1.00	0.83	864
accuracy			0.85	2237
macro avg	0.86	0.88	0.85	2237
weighted avg	0.89	0.85	0.85	2237

(c)

Fig. 17 Classification report concerning **a** Image-only model [VIT Base 16], **b** Text-only model (GPT-2), and **c** Image + text model (random forest classifier model)

be extended with some advanced studies by implementing new age transformer models such as T5 [51] for text classification and CoAtNet [52] for image classification. Since our models for each modality are trained independently, a combination of different architectures can be used to analyze the result and performance of each model on different datasets. In addition, a fusion of models can be tried by using other ensemble techniques like stacking to obtain better prediction results.

Besides, the imbalance in text data can be removed by filtering it on the basis of the labels. Since this study only covers a portion of the dataset about one task; future work can include other dataset tasks to examine the performance of transformer models on them. We have not performed any filtration on the data to use the whole dataset resulting in a slightly imbalanced dataset for text modality. Finding an appropriate approach to handle this problem by developing a system that predicts the informativeness of crisis-related tweets more effectively can be considered a future improvement. Forecasting models can also be effective in this aspect [53–55].

Future work could include comparisons with alternative encoders, such as CNN-based models for images or BERT for text, to evaluate performance variations. While this study does not exhaustively compare all encoding options, the innovation lies in the novel integration mechanism that enhances classification accuracy. One limitation of this study is to make use of a single dataset, which may restrict the generalizability of the findings to other domains. Additionally, while ViT and GPT-2 were effective in our setup, their computational demands may pose challenges for deployment in resource-constrained environments. The choice to use a single dataset was due to time and resource constraints. However, we ensured diverse coverage within the dataset to partially address this concern and to make it robust. Future research could focus on evaluating the proposed approach across multiple datasets from diverse domains to improve its generalizability. Additionally, exploring lightweight encoder models for reducing computational complexity could enhance the practicality of the method for real-time applications.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions. This research was funded by the Symbiosis International (Deemed University), Pune, India, and the Centre for Advanced Modelling & Geospatial Information Systems (CAMGIS), Faculty of Engineering & IT, University of Technology Sydney. Also, supported by the Ongoing Research Funding program (ORF-2025-14), King Saud University, Riyadh, Saudi Arabia.

Code Availability The authors have made the code publicly available via the link below: https://github.com/Arundarasi89/Multi-modal-codes.

References

1. Lahat, D.; Adali, T.; Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. Proc. IEEE **103**(9), 1449–1477 (2015). https://doi.org/10.1109/JPROC.2015.2460697
2. Giannakos, M.N.; Sharma, K.; Pappas, I.O.; Kostakos, V.; Velloso, E.: Multimodal data as a means to understand the learning experience. Int. J. Inf. Manag. **48**, 108–119 (2019). https://doi.org/10.1016/j.ijinfomgt.2019.02.003
3. Liu, X.; Wang, M.; Huet, B.: Event analysis in social multimedia: a survey. Front. Comput. Sci. **10**(3), 433–446 (2016). https://doi.org/10.1007/s11704-015-4583-2
4. Gautam, A.K., Misra, L., Kumar, A., Misra, K., Aggarwal, S., Shah, R.R.: Multimodal analysis of disaster tweets. In: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), pp. 94–103. IEEE (2019). https://doi.org/10.1109/BigMM.2019.00-38
5. Duong, C. T., Lebret, R., Aberer, K.: Multimodal classification for analysing social media (2017). arXiv preprint http://arxiv.org/abs/1708.02099. https://doi.org/10.48550/arXiv.1708.02099
6. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 569–576 (2017)
7. Mouzannar, H., Rizk, Y., Awad, M.: Damage Identification in Social Media Posts using Multimodal Deep Learning. In: ISCRAM (2018)
8. Ofli, F., Alam, F., Imran, M.: Analysis of social media data using multimodal deep learning for disaster response (2020). arXiv preprint http://arxiv.org/abs/2004.11838. https://doi.org/10.48550/arXiv.2004.11838
9. Gallo, I., Ria, G., Landro, N., La Grassa, R.: Image and text fusion for UPMC Food-101 using BERT and CNNs. In: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), pp. 1–6. IEEE (2020). https://doi.org/10.1109/IVCNZ51579.2020.9290622.
10. Zou, Z.; Gan, H.; Huang, Q.; Cai, T.; Cao, K.: Disaster image classification by fusing multimodal social media data. ISPRS Int. J. Geo Inf. **10**(10), 636 (2021). https://doi.org/10.3390/ijgi10100636
11. Gallo, I., Calefati, A., Nawaz, S.: Multimodal classification fusion in real-world scenarios. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 5, pp. 36–41. IEEE (2017). https://doi.org/10.1109/ICDAR.2017.326
12. Velioglu, R., Rose, J.: Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge (2020). arXiv preprint http://arxiv.org/abs/2012.12975. https://doi.org/10.48550/arXiv.2012.12975.
13. Audebert, N., Herold, C., Slimani, K., Vidal, C.: Multimodal deep networks for text and image-based document classification. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 427–443. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-43823-4_35
14. Miller, S.J.; Howard, J.; Adams, P.; Schwan, M.; Slater, R.: Multimodal classification using images and text. SMU Data Sci. Rev. **3**(3), 6 (2020)
15. Zhang, K.; Geng, Y.; Zhao, J.; Liu, J.; Li, W.: Sentiment analysis of social media via multimodal feature fusion. Symmetry **12**(12), 2010 (2020). https://doi.org/10.3390/sym12122010
16. Alam, F., Ofli, F., Imran, M.: Crisismmd: multimodal twitter datasets from natural disasters. In: Twelfth International AAAI Conference on Web and Social Media (2018)
17. "CrisisMMD: Multimodal Crisis Dataset", Crisis NLP, https://crisisnlp.qcri.org/crisismmd, Accessed 27 April 2022
18. Bostrom, K., Durrett, G.: Byte pair encoding is suboptimal for language model pretraining (2020). arXiv preprint http://arxiv.org/abs/2004.03720. https://doi.org/10.48550/arXiv.2004.03720
19. Park, K., Lee, J., Jang, S., Jung, D.: An empirical study of tokenization strategies for various Korean NLP tasks (2020). arXiv preprint http://arxiv.org/abs/2010.02534. https://doi.org/10.48550/arXiv.2010.02534
20. ALEnezi, N.S.A.: A method of skin disease detection using image processing and machine learning. Procedia Comput. Sci. **163**, 85–92 (2019). https://doi.org/10.1016/j.procs.2019.12.090
21. Khalifa, N.E.; Loey, M.; Mirjalili, S.: A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif. Intell. Rev. (2021). https://doi.org/10.1007/s10462-021-10066-4
22. Elgendi, M.; Nasir, M.U.; Tang, Q.; Smith, D.; Grenier, J.P.; Batte, C.; Nicolaou, S.: The effectiveness of image augmentation in deep learning networks for detecting COVID-19: a geometric transformation perspective. Front. Med. (2021). https://doi.org/10.3389/fmed.2021.629134
23. Yan, J., Lin, S., Bing Kang, S., Tang, X.: Learning the change for automatic image cropping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 971–978 (2013)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhin, I.: Attention is all you need. Adv. Neural Inf. Process. Syst. **30** (2017)
25. Ott, M., Edunov, S., Grangier, D., Auli, M.: Scaling neural machine translation. In: WMT (2018)
26. Khan, S.; Naseer, M.; Hayat, M.; Zamir, S.W.; Khan, F.S.; Shah, M.: Transformers in vision: a survey. ACM Comput. Surv. (CSUR) (2021). https://doi.org/10.1145/3505244
27. Chaudhari, S.; Mithal, V.; Polatkan, G.; Ramanath, R.: An attentive survey of attention models. ACM Trans. Intell. Syst. Technol. (TIST) **12**(5), 1–32 (2021). https://doi.org/10.1145/3465055
28. Correia, A.D.S., Colombini, E.L.: Attention, please! A survey of neural attention models in deep learning (2021). arXiv preprint http://arxiv.org/abs/2103.16775. https://doi.org/10.48550/arXiv.2103.16775
29. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding (2018). arXiv preprint http://arxiv.org/abs/1810.04805
30. Bengio, Y.; Goodfellow, I.; Courville, A.: Deep Learning, Vol. 1. MIT Press, Cambridge (2017)
31. Hochreiter, S.; Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
32. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A.: Vision transformers for remote sensing image classification. Remote Sens. **13**(3), 516 (2021). https://doi.org/10.3390/rs13030516
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
34. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

35. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. **25** (2012)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint http://arxiv.org/abs/1409.1556. https://doi.org/10.48550/arXiv.1409.155
37. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2014). arXiv preprint http://arxiv.org/abs/1409.1556. https://doi.org/10.48550/arXiv.1409.4842
38. Bai, Y., Mei, J., Yuille, A.L., Xie, C.: Are transformers more robust than CNNs? Adv. Neural Inf. Process. Syst. **34**
39. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale (2020). arXiv preprint http://arxiv.org/abs/2010.11929. https://doi.org/10.48550/arXiv.2010.11929
40. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness (2018). arXiv preprint http://arxiv.org/abs/1811.12231. https://doi.org/10.48550/arXiv.1811.12231
41. Goodfellow, I. J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples (2014). arXiv preprint http://arxiv.org/abs/1412.6572. https://doi.org/10.48550/arXiv.1412.6572
42. Nhu, V.H.; Shirzadi, A.; Shahabi, H.; Chen, W.; Clague, J.J.; Geertsema, M.; Lee, S.: Shallow landslide susceptibility mapping by random forest base classifier and its ensembles in a semi-arid region of Iran. Forests **11**(4), 421 (2020). https://doi.org/10.3390/f11040421
43. Feng, Z., Mo, L., Li, M.: A Random Forest-based ensemble method for activity recognition. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5074–5077. IEEE (2015). https://doi.org/10.1109/EMBC.2015.7319532
44. Shaik, A.B., Srinivasan, S.: A brief survey on random forest ensembles in classification model. In: International Conference on Innovative Computing and Communications, pp. 253–260. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-2354-6_27
45. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9650–9660 (2021). https://doi.org/10.48550/arXiv.2104.14294
46. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Rush, A. M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. https://doi.org/10.18653/v1/2020.emnlpdemos.6
47. Sanyal, D., Bosch, N., Paquette, L.: Feature Selection Metrics: Similarities, Differences, and Characteristics of the Selected Models. International Educational Data Mining Society (2020)
48. Hajian-Tilaki, K.: Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J. Intern. Med. **4**(2), 627 (2013)
49. Popel, M., Bojar, O.: Training tips for the transformer model (2018). arXiv preprint http://arxiv.org/abs/1804.00247. https://doi.org/10.48550/arXiv.1804.00247.
50. Madichetty, S.; Muthukumarasamy, S.; Jayadev, P.: Multi-modal classification of Twitter data during disasters for humanitarian response. J. Ambient. Intell. Humaniz. Comput. **12**(11), 10223–10237 (2021). https://doi.org/10.1007/s12652-020-02791-5
51. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2019). arXiv preprint http://arxiv.org/abs/1910.10683
52. Dai, Z.; Liu, H.; Le, Q.V.; Tan, M.: Coatnet: marrying convolution and attention for all data sizes. Adv. Neural Inf. Process. Syst. **34**, 3965–3977 (2021)
53. Htun, H.H.; Biehl, M.; Petkov, N.: Forecasting relative returns for S&P 500 stocks using machine learning. Financ. Innov. **10**, 118 (2024). https://doi.org/10.1186/s40854-024-00644-0
54. Cheng, L.C.; Lu, W.T.; Yeo, B.: Predicting abnormal trading behavior from internet rumor propagation: a machine learning approach. Financ. Innov. **9**, 3 (2023). https://doi.org/10.1186/s40854-022-00423-9
55. Antonio, B.O.; Juan, L.R.; Ana, I.D., et al.: Examining user behavior with machine learning for effective mobile peer-to-peer payment adoption. Financ. Innov. **10**, 94 (2024). https://doi.org/10.1186/s40854-024-00625-3