IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 August 2022, accepted 22 August 2022, date of publication 29 August 2022, date of current version 9 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3202976

RESEARCH ARTICLE

CAMM: Cross-Attention Multimodal Classification of Disaster-Related Tweets

ANURADHA KHATTAR[1,2] **AND S. M. K. QUADRI**[1]
[1]Department of Computer Science, Jamia Millia Islamia, New Delhi 110025, India
[2]Department of Computer Science, Miranda House, University of Delhi, New Delhi 110007, India

Corresponding author: Anuradha Khattar (anuradha.khattar@mirandahouse.ac.in)

ABSTRACT During the past decade, social media platforms have been extensively used for information dissemination by the affected community and humanitarian agencies during a disaster. Although many studies have been done recently to classify the informative and non-informative messages from social media posts, most are unimodal, i.e., have independently used textual or visual data to build deep learning models. In the present study, we integrate the complementary information provided by the text and image messages about the same event posted by the affected community on the social media platform Twitter and build a multimodal deep learning model based on the concept of the attention mechanism. The attention mechanism is a recent breakthrough that has revolutionized the field of deep learning. Just as humans pay more attention to a specific part of the text or image, ignoring the rest, neural networks can also be trained to concentrate on more relevant features through the attention mechanism. We propose a novel Cross-Attention Multi-Modal (CAMM) deep neural network for classifying multimodal disaster data, which uses the attention mask of the textual modality to highlight the features of the visual modality. We compare CAMM with unimodal models and the most popular bilinear multimodal models, MUTAN and BLOCK, generally used for visual question answering. CAMM achieves an average F1-score of 84.08%, better than the MUTAN and BLOCK methods by 6.31% and 5.91%, respectively. The proposed cross-attention-based multimodal deep learning method outperforms the current state-of-the-art fusion methods on the benchmark multimodal disaster dataset by highlighting more relevant cross-domain features of text and image tweets.

INDEX TERMS Deep convolutional neural network (DCNN), disaster management, multimodal learning, attention mechanism, cross-attention, social media, Twitter.

I. INTRODUCTION

In the past decade, emergency managers and safety organizations have started using social media platforms to share critical information for planning and implementing rescue operations during a disaster. The decision-makers utilize timely, first-hand, and location-based messages posted by eyewitnesses on social media platforms to deploy resources and enhance their response efforts. Innovative use of these platforms allows humanitarian teams to engage directly with the affected public during all phases of disaster management. Among several social media platforms, Twitter is most prevalent during natural disasters [1], [2]. Twitter text messages, called tweets that consist of up to 280 characters, give first-hand information about the event almost in real time. A massive flood of tweets is generated on Twitter within minutes of striking a disaster [3], [4], [5], [6]. With advancing mobile technologies, text tweets are often accompanied by related images or videos, providing complementary information to understand the disaster site situation better. Analyzing these multimodal posts together for an event allows the government authorities and humanitarian organizations to assess the post-disaster situation from different angles and perspectives to take appropriate action. While these tweets provide crucial information during an emergency, filtering informative and actionable messages from a vast pool of these noisy messages is challenging [7], [8].

Although several artificial intelligence-based tools have been proposed recently to make sense of this enormous crisis data and filter out relevant messages, most of these methods

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Tan.

are based on a single modality. For example, these methods have independently used text [9], [10], [11], [12], [13], images [14], [15], [16], or videos [17] posted on social media platforms but have not fully explored recent multimodal techniques to exploit the complementary information provided by more than one modality.

In the recent past, deep multimodal learning has gained vast popularity. It is being applied to several fields like visual question answering [18], [19], sentiment analysis [20], [21], neural machine translation [22], cross-modal retrieval [23], speech recognition [24] and image captioning. However, very few studies have used the concept of deep multimodal learning to classify Twitter messages (text, images, and videos) posted during an ongoing disaster. The present study proposes a cross-attention-based deep multimodal learning method that uses text and image modalities of disaster tweets.

Under the multimodal category, the most common methods used by researchers to fuse the features of text and image modalities for the classification of disaster-related data are early fusion and late fusion [25], [26], [27], [28], [29]. *Early fusion* is also called feature-based fusion, where the final decision is based on the common vector obtained after concatenating the extracted features of the individual modalities. This method takes advantage of learning from the low-level interactions of features of all the modalities. On the other hand, in the case of *late fusion*, also known as decision fusion, the decision for each modality is made, and then these decisions are combined to get the final decision [30].

One recent breakthrough that has revolutionized the field of deep learning is the attention mechanism [31]. Just like humans pay more attention to a specific part of the text or image, ignoring the rest, the neural networks can also be trained to concentrate on more relevant features through the attention mechanism. Initially, attention was more prevalent in Natural Language Processing [32], but recently it is also applied to images, videos, and audio modalities. The newly introduced attention mechanism focused on some specific regions of the feature map achieves better performance. If the attention mask from any modality highlights the features in the same modality, it is called *self-attention*. In contrast, if the attention mask of one modality is used to highlight the features of another modality, it is called *cross-attention*. Cross-attention model not only learns the complementary features of the two modalities but is also able to filter the noise to give better results. In the present study, we propose a cross-attention-based classification method that uses the attention mask of text tweets to highlight the image tweet's features.

The limitation of existing early and late fusion methods is that they assign a fixed weight to each modality which is overcome by the proposed attention-based method that balances the contribution of modalities by dynamically assigning weights to the features of different modalities. This allows the attention model to choose relevant, more prominent, and complimentary features from each modality. Our motivation for using the multimodal approach is to explore the relationship between the two media and use them harmoniously to achieve better results. The only constraint of the attention-based method is the additional computation of the attention weights that is outdone by the improved network performance.

Based on the above discussion, the main contributions of the present study are:

- We propose a deep multimodal network designed to learn the prominent features from the textual and visual modalities using a novel **C**ross-**A**ttention **M**ulti-**M**odal (CAMM) framework for the binary classification of disaster tweets into '*informative*' and '*non_informative*' classes. CAMM is designed to utilize the complementary information from the tweets' textual and imaging modalities. The attention mask of text modality is used to highlight the features of the imaging modality. Our goal is to attenuate the image features by determining the relationship between the words in the tweet and different spatial regions in an image. The multimodal dataset used in this study is CrisisMMD [33], comprising the text and image tweets of seven disasters posted on Twitter during 2017.
- To validate the performance of the proposed model CAMM, we perform experiments under the following setups, which serve as baselines:
  (i) Unimodal classification of text tweets using Bidirectional Long Short-Term Memory (Bi-LSTM).
  (ii) Unimodal classification of image tweets using DCNN.
  (iii) Multimodal classification using MUTAN [34] fusion method.
  (iv) Multimodal classification using BLOCK [35] fusion method.
- We also compare the results of CAMM with five recent state-of-the-art multimodal disaster-related studies based on CrisisMMD dataset [25], [26], [27], [28], [29].

To the best of our knowledge, a cross-attention-based, multimodal fusion approach has not yet been explored in the context of social media disaster data classification.

The rest of the paper is structured as follows: In Section II, we discuss the research work related to unimodal and multimodal techniques for disaster management proposed in the recent past. Section III covers the architecture of the proposed deep multimodal neural network CAMM. A brief overview of two baseline multimodal models is given in section IV. The experimental setup in Section V includes the dataset, metrics, hyperparameters, and the baseline methods used for performing the experiments. The implementation details of the experiments performed under various setups are given in section VI. We list and discuss the results obtained after training the networks under five different setups in section VII. Finally, in section VIII, we discuss the limitations and future scope of the work.

II. RELATED WORK

Recently with the popularity of social media platforms at the time of emergency, messages comprising situational information, warnings, sentiments, infrastructural damage, geographical information, or medical help are posted on these platforms in abundance [36]. These messages may be in the form of text, images, or videos. This massive volume of rich multimodal data can be converted into useful information by scientists and domain experts to help the response team in rescue operations and control the impact of the disaster [37]. Although several unimodal models based on text and images have been proposed in the last decade, the work on multimodal modalities is relatively recent. This section summarizes the research work done in handling social media disaster data under three categories: (i) Unimodal methods based on text-only modality, (ii) Unimodal methods based on image-only modality, and (iii) Multimodal techniques based on more than one modality proposed in the recent past.

A. UNIMODAL: BASED ON TEXT-ONLY MODALITY

Many Natural Language Processing (NLP) based methods have been applied recently to the text messages posted on various social media platforms [38]. For example, Rudra *et al.* [12] filtered out the situational awareness messages from Twitter by identifying tweets' low-level features and summarized these real-time streams of tweets. Basu *et al.* [39] worked on the Nepal earthquake tweets to match the need-based tweets with the supply-based tweets. Similarly, Purohit *et al.* [9] attempted to prioritize the requests made by the affected people to be serviced by the emergency responders. In a recent study by Madichetty *et al.* [40], authors filtered 'Need and Availability Resources (NAR)' tweets from the Nepal and Italy earthquakes that happened in 2015 and 2016, respectively, using a stacked architecture of CNN with traditional classifiers. Their experiments on various classifiers confirm that K-Nearest-Neighbor as the base classifier, Support Vector Machine as the meta classifier, and CNN give the best results. Suwaleih *et al.* [41] emphasized the mention of the location or place in the tweet message posted during a disaster. They compared the performance of the 'Location Mention Recognition (LMR)' task on crisis-related and general datasets consisting of text tweets. The authors fine-tuned the pre-trained BERT model for training on five datasets. Their results confirm that crisis-related tweets from locations near the disaster event are most helpful for the first responders. Sufi *et al.* [42] presented an NLP-based system to understand location-oriented sentiments on the most extensive set of languages. Their system showed an accuracy of 97% when tested on the live feed of 67515 tweets. In another study by Zahera *et al.* [43] on Text REtrieval Conference-Incident Stream (TREC-IS) [44] dataset and COVID-19 tweets, the authors classified the tweets into multiple categories where each tweet may belong to more than one category. They used three models in their study, (i) BERT for tweet vectorization, (ii) graph attention network (GAT) to understand the relation between the tweets and labels, and (iii) proposed a metric to compute the distance between the vectors produced by (i) and (ii). Their model achieved an average F1-score of 59% on TREC-IS and 55% on COVID-19 datasets. Since labeling is a tedious and expensive task, self-labeling [45], [46], synthetic labeling [47], and semi-supervised learning [48] methods have also been proposed recently. At the onset of a disaster, the unavailability of labeled data has also encouraged researchers to propose methods based on transfer learning and domain adaptation. Li *et al.* were among the first few researchers to explore this area. They used an iterative self-training strategy using soft and hard labels to identify relevant tweets and labeled data of earlier disasters to learn the classifier for the current disaster [49], [50]. Imran *et al.* also proposed very effective models based on convolutional neural networks and domain adaptation for disaster management [13], [51].

B. UNIMODAL: BASED ON IMAGE-ONLY MODALITY

In a recent study by Ahadzadeh and Mohammad [52], the machine learning methods Support Vector Machine, and Naïve Bayes are applied to tweet images to assess the damage done due to earthquakes. Studies by Khattar and Quadri compared the simple transfer learning, unsupervised domain adaptation, and semi-supervised domain adaptation approaches applied to the natural and biological disaster image datasets [16], [53]. Robertson *et al.* [54] fine-tuned pre-trained model VGG-16 on Hurricane Harvey images to classify them on an 'urgency' and 'time-period' basis. In a similar study, Li *et al.* [15] applied Domain Adversarial Neural Network (DANN) on four disaster images for binary classification into 'Damage' and 'No-damage'.

C. MULTIMODAL: BASED ON BOTH TEXT AND IMAGE MODALITIES

Under the multimodal analysis, Gautam *et al.* [25] proposed a diffusion method for the classification of Twitter data (text and images) of seven disasters of the CrisisMMD dataset [33] into two classes 'informative' and 'non-informative' and compared their model with the unimodal models based on text-only and image-only modalities. For text-only modality, they applied N-gram, LSTM, BiLSTM, and CNN+Glove methods, and for image-only modality, they used six pre-trained models VGG-16, VGG-19, ResNet50, InceptionV2, Xception, and DenseNet for transfer learning. Finally, they compared the results based on three Logistic Regression Decision policies. Their results confirm that the logistic regression decision policy with bigram for text and ResNet50 for images gives the best results.

Authors Ofli *et al.* [26] emphasized that complementary information from different modalities leads to more robust inference. They worked on the same CrisisMMD [33] dataset as Gautam *et al.* with filtered records where the text label is the same as the image label. High-level features are extracted from two parallel networks, one for text messages and another for images. They used CNN with five hidden layers for the tweet text messages, and for the images, they used a

pre-trained network VGG-16. The feature vectors from these two networks are combined and passed to a dense layer followed by SoftMax. This type of feature handling from two modalities comes under the category of early fusion. Their experiments showed that the multimodal early fusion-based model performs better than the unimodal models for data classification, first into informative / non-informative and then into five humanitarian categories.

The work of Li *et al.* [27] is also based on the CrisisMMD dataset, and they performed their multimodal experiments under two scenarios: (a) text and image labels do not match (b) text and image labels match. For the unimodal text model, they used RNNs Gated Recurrent Unit (GRU) variant, and for images, they used CNN. First, they trained two networks for text and images independently and then took the average of each network prediction as the final output. In the second case, they concatenated the feature vectors of the text and image networks and passed the combined vector to the dense layer to get the output. They concluded that the cancat+train (late fusion) method performs better than the average (early fusion) method.

In a similar study, Madichetty *et al.* [28] applied the concept of late fusion by adding the feature vectors of the two modality networks to predict the final label. Text tweets are trained on ANN, and VGG-16 is used as the pre-trained network for the images.

Kumar *et al.* [29] applied the majority voting scheme to the predicted classes of unimodal models for 'text' (LSTM) and 'image' (VGG-16) and multimodal model for 'text + image'. For the multimodal classification, they used late fusion. The final label assigned to the class is the label with two or more votes.

In another study, Pouyanfar *et al.* [17] classified the videos of two hurricanes, Harvey and Maria, using a new multimodal classification deep learning framework. For the audio component, SoundNet, a pre-trained model, was applied. For the visual-spatial component InceptionV3 and for the temporal-video component, LSTM is followed by a dense layer. For the final scores, they proposed to combine the audio and visual scores with a Multiple Component Analysis based fusion model that showed the effectiveness of the proposed model over other existing techniques. Nie *et al.* [55] used three modalities (point cloud, Multiview, and panorama view) to represent a 3D shape and proposed two novel loss functions: correlation loss and instance loss. They used a weighted method to fuse these modalities to build a robust model for 3D shape recognition.

From the above discussion, it is clear that researchers have shown greater interest recently in exploring the field of deep multimodal learning for disaster-related research. Most of these studies have applied feature-based (early) and decision-based (late) fusion methods to combine the features of multiple modalities. However, the recently introduced attention mechanism for multimodal learning is not fully explored for disaster management, which motivated us to fill this research gap.

III. PROPOSED WORK

We propose a novel architecture to build a binary classifier that integrates the information about the same event expressed in two different ways in the form of words and pictures. Fig. 1 shows the complete architecture of the proposed multimodal DCNN that uses annotated text and image tweets posted on Twitter during seven disasters.

We are given a tweet text $\boldsymbol{T}$ and a tweet image $\boldsymbol{I}$, and we need to fuse the features of $\boldsymbol{T}$ and $\boldsymbol{I}$ to predict the final class as '*informative*' or ' *non_informative*'. As commonly done in multimodal architectures, the text $\boldsymbol{T}$ and image $\boldsymbol{I}$ are first converted to vector representations. Then these representations are fused to extract the most meaningful interactions between the text and the image to get the predicted class. In this study, we propose a new fusion technique called Cross-Attention Multi-Modal (CAMM) fusion, where the features extracted for each word in a tweet ''attend'' to different spatial regions of the image features. Our motivation for this approach is that different words in a tweet can accentuate relevant image features that significantly improve the model's performance.

As shown in Fig. 1, we use the pre-trained model VGG-16 to extract features from the input image $\boldsymbol{I}$. The output of convolution layers is passed through the Tanh() activation function to limit the range of features between -1 and 1. For a tweet $\boldsymbol{T}$ with n words, we use a Bi-LSTM with two layers to learn hidden representations of dimension d_T for each word. Finally, we represent the image and text features by F_I and F_T, respectively.

In self-attention, a feature vector is generated for each word in the string, then the three weight matrices W_K, W_Q and W_V are used on the words to extract the key, query, and value vectors for each word. In the proposed cross-attention structure, our goal is to attenuate the image features by determining the relationship between the words in the tweet and different spatial regions in an image. The W_K matrix is used to extract key values for each word in the tweet, and W_Q is used on feature vector of an image obtained using a CNN. The key vectors obtained for each word and query vectors obtained for each spatial region in an image are then combined to create the attention given by the following equation,

$$A = softmax\left(\frac{W_Q F_I \times (W_K F_T)^{\mathrm{T}}}{\sqrt{d_k}}\right), \tag{1}$$

where d_k represents the dimension of the output of W_K F_T. This attention map represents the effect of each word in the tweet on different spatial regions in the image. Finally, we generate the value vectors given by, $W_V F_I$, which are then multiplied with the attention map A to generate the ''attended'' image features,

$$F_{CA} = A \odot (W_V F_I), \tag{2}$$

where $\odot$ represents element-wise multiplication. The attended image features are subsequently flattened and passed through the classification layers to predict whether the image-text pair is '*informative*' or '*non_informative*'.

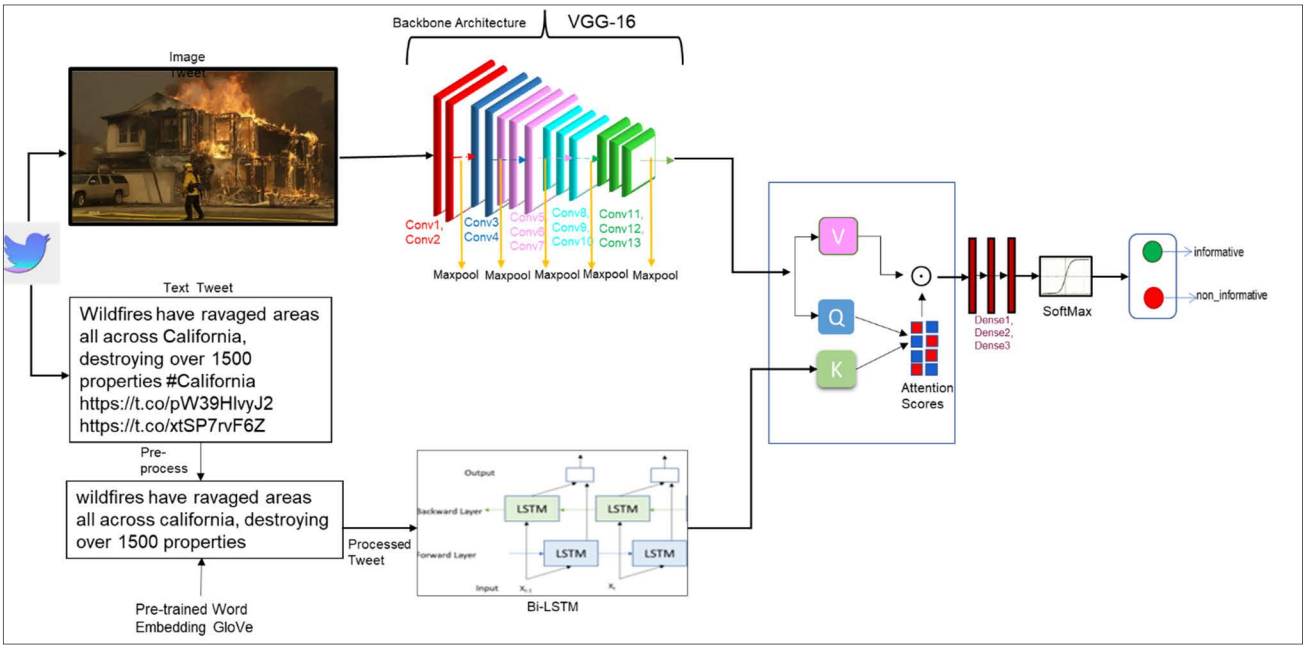

FIGURE 1. The architecture of the proposed Cross-Attention Multi-Modal (CAMM) model for binary classification of disaster-related tweets.

IV. A BRIEF OVERVIEW OF BASELINE MULTIMODAL MODELS

We train two popular multimodal models, MUTAN [34] and BLOCK [35], proposed by Younes *et al.* on the CrisisMMD dataset as baseline models and compare their results with the proposed model, CAMM. We briefly discuss these models in this section.

MUTAN is a multimodal fusion method that aims to capture the relation between every pair of neurons of text and image modalities. The fused features can be represented by the relation:

$$y_k = \sum_{i=1}^{d_T} \sum_{j=1}^{d_I} M_{ijk} F_T^i F_I^j \tag{3}$$

$$y = M \times_1 F_T \times_2 F_I \tag{4}$$

where y_k is the output at the kth index
F_T is the text vector with dimension d_T
F_I is the image vector with dimension d_I
M is the 3D learnable weight matrix
y is the output

MUTAN applies Tucker decomposition to reduce the number of parameters of the weight matrix, M. M is represented as a product of W_T, W_I, W_y and a core tensor M_c as:

$$y_{out} = W_y \left(M_c \times_1 (W_T F_T) \times_2 (W_I F_I)\right) \tag{5}$$

where W_T, W_I and W_y are the weight matrices to reduce the dimension of the text features, image features, and the fused vector.

As explained above, MUTAN compresses the feature vectors of text (F_T) and image (F_I) to reduce the number of parameters in the fused vector. However, this may lead to loss of information if the reduced-sized feature vectors do not efficiently capture all the information in the text and image features. Younes *et al.* [35] proposed the BLOCK fusion model to address this problem. This method divides the feature vectors, F_T and F_I into blocks and applies Tucker decomposition to each individual block. It finally concatenates them to obtain the final fused feature vector.

V. EXPERIMENTAL SETUP

Extensive experiments are conducted on the multimodal dataset of seven disasters for the proposed cross-attention-based multimodal framework. The results are compared with text-only, image-only, and two multimodal baseline models. All experiments are performed three times, and the average is reported.

A. DATASET

The benchmark multimodal disaster dataset used in the present study is the 'CrisisMMD' Twitter dataset released by Imran *et al.* [33]. This dataset has images and text tweets posted for seven devastating natural disasters held in 2017 worldwide. These disasters were: "Hurricane Harvey, Hurricane Irma, Hurricane Maria, Mexico Earthquake, California Wildfires, Iraq-Iran earthquake, and Sri Lanka Floods". The tweets are filtered and manually annotated into two classes: {'*informative*', '*non_informative*'}. *'informative'* tweets provide actionable information for the humanitarian agencies about dead, injured, or lost people and infrastructural damage. Advice, warnings, and requests for aid also come in this category. For example, Fig. 2 shows a batch of informative images and text tweets posted during Hurricane Harvey and the Sri Lanka floods. Informative

TABLE 1. CrisisMMD Dataset [33] with matching labels for text and image tweets.

Disaster	Duration of tweet collection	Total Tweets	Informative	Non_informative	Train Set	Val Set	Test Set
Hurricane Harvey	26-Aug-17 to 20-Sep-17	3168	2262	906	2027	507	634
Hurricane Irma	06-Sep-17 to 21-Sep-17	2799	2032	767	1791	448	560
Hurricane Maria	20-Sep-17 to 13-Nov-17	3108	1813	1295	1988	498	622
Mexico Earthquake	20-Sep-17 to 06-Oct-17	1121	806	315	716	180	225
California Wildfires	10-Oct-17 to 27-Oct-17	1205	923	282	771	193	241
Iraq Iran Earthquake	13-Nov-17 to 19-Nov-17	500	398	102	320	80	100
Sri Lanka Floods	31-May17 to 03-Jul-17	861	229	632	550	138	173

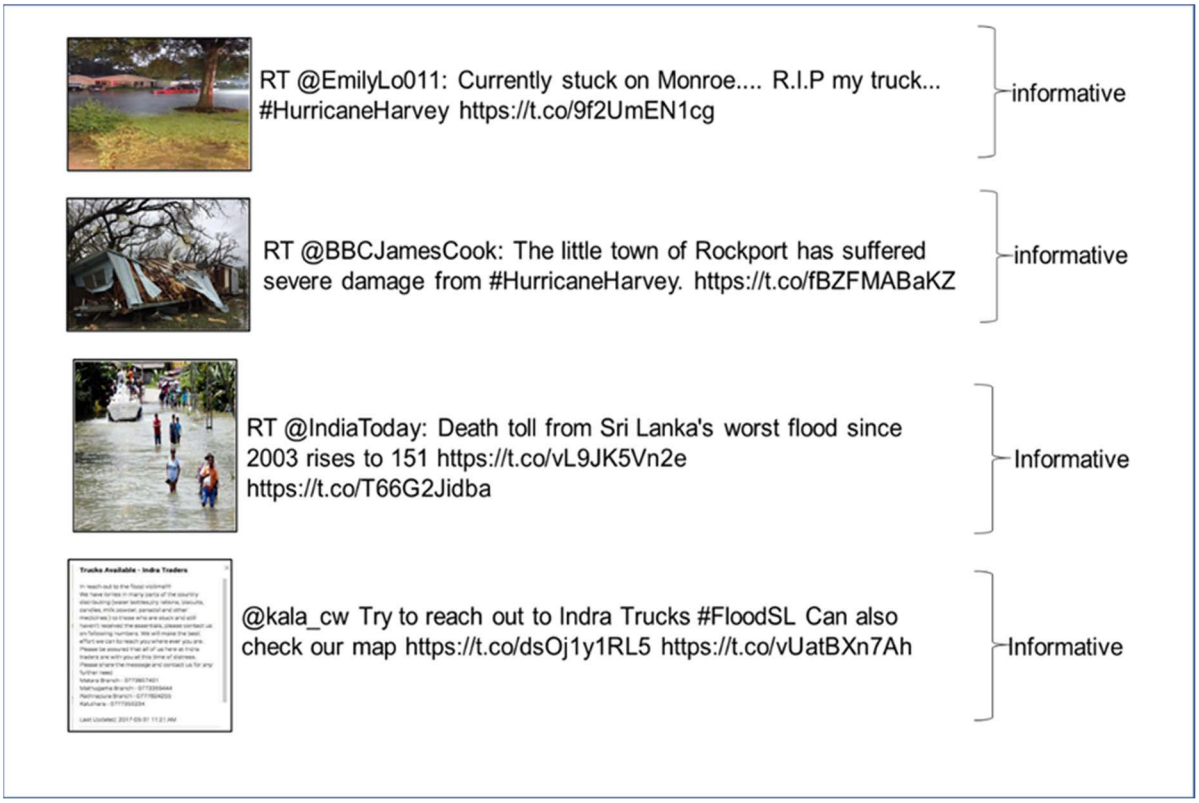

FIGURE 2. Informative image and text tweets posted during Hurricane Harvey and Sri Lanka Floods. These types of messages.

messages are beneficial for humanitarian agencies in rescue operations.

Fig. 3 shows a batch of '*non_informative*' tweets which may include text messages expressing sympathy for the affected people or may comprise opinions in general. Images of politicians visiting the disaster site, posters, or logos that do not provide helpful information in disaster response also belong to the '*non_informative*' category.

The dataset consists of 16058 text tweets and 18082 image tweets, as up to four pictures can be posted along with one text tweet on Twitter. However, since the text and the corresponding images are annotated separately, their labels may not be aligned. are very helpful for the humanitarian agencies in the rescue operations.

We have filtered only those tweets with the same label for the text and the corresponding image for the present study. Table 1 gives the details of the filtered dataset with 12762 tweets each for the image and the text modalities, out of which 8463 are informative, and 4299 are non-informative. The filtered dataset is further split into the train, val, and test set in the ratio 80:10:10 for training, validation, and testing purpose. Most disaster-related studies done in the recent past based on multimodal data analysis have used the CrisisMMD dataset.

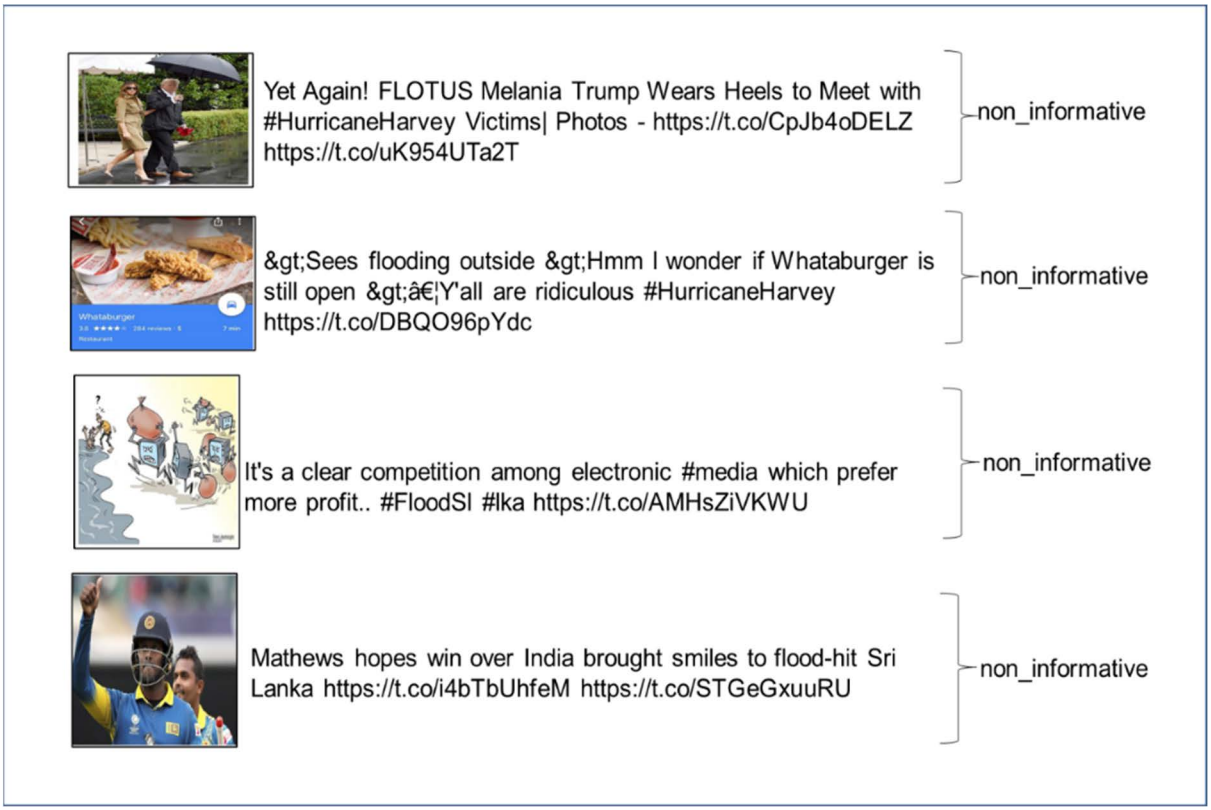

FIGURE 3. Non_informative image and text tweets posted during Hurricane Harvey and Sri Lanka Floods. These include posters, banners or other messages that do not provide any useful information.

As our dataset is imbalanced, the accuracy metric is unsuitable for evaluating the model's performance. Instead, we compute weighted F1-score and AUC metrics to handle this issue. Also, we use weighted cross-entropy loss, where the weight assigned for each class is inversely proportional to the number of class samples in the training set.

B. METRICS USED

To estimate the performance of the trained models and to do a comparative analysis we have used the following metrics in the present study:

(a) Accuracy: Ratio of correct predictions made by the model and the total samples.

$$\text{Accuracy} = \frac{\text{True\_Positive} + \text{True\_Negative}}{\text{Total}}$$

(b) Precision: Ratio of correctly predicted positive samples and all the predicted positive samples.

$$\text{Precision} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Positive}}$$

(c) Recall: Ratio of correctly predicted positive samples and the actual positive samples.

$$\text{Recall} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Negative}}$$

(d) F1-score: The F1-score is computed by taking the Harmonic Mean (HM) of precision and recall.

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

(e) ROC-AUC: ROC curve stands for Receiver Operator Characteristic curve drawn between True Positive Rate (TPR) on the y-axis and False Positive Rate (FPR) on the x-axis at various thresholds and is generally used for binary classification. Area Under the Curve (AUC) summarizes the ROC curve, a higher value of AUC results in a better performing model.

C. HYPERPARAMETER SETTING

All the experiments are performed under the PyTorch framework on Tesla P100 GPU with High-RAM provided by Google Colab Pro. We have applied the Grid Search method to finetune the hyperparameters. We choose various combinations of the hyperparameters and compare the model's performance for every combination to select the optimal value. Although this method is relatively slow, it helps find the best values for the network parameters. The parameters selected through grid search for training all the models include: learning rate as 1.00e-03, weight decay as 5.00e-04, momentum as 0.9, loss function as weighted CrossEntrpoyLoss and optimizer as Stochastic Gradient Descent (SGD). We performed 50 epochs for the baseline models and 100 epochs for the

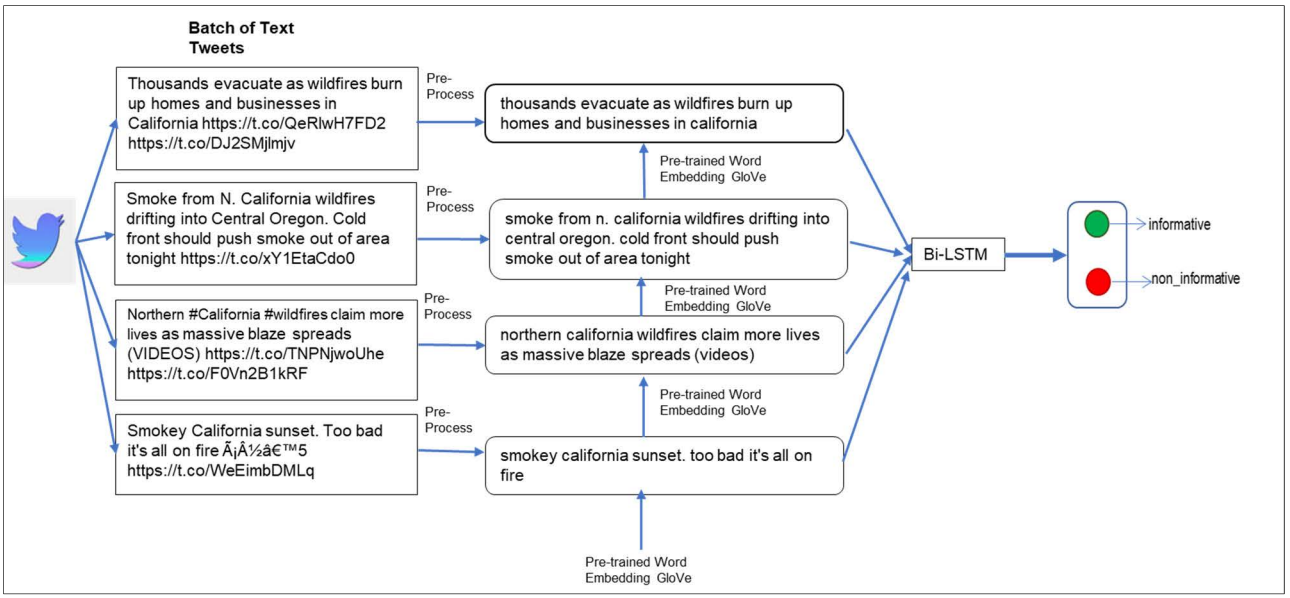

FIGURE 4. Unimodal classification of text tweets using Bi-LSTM and pre-trained word embedding GloVe.

proposed CAMM model. The learning rate schedular is set at [25], [35] for 50 epochs and at [60, 75] for 100 epochs. For DCNN, we have taken VGG-16 as the backbone architecture. For text, Bi-LSTM uses two layers which are enough to capture the complex features of the text tweet. An Ablation study for the choice of hyperparameters is discussed in section VII C, and the results are listed in Tables 2, 3, and 4.

D. BASELINES

We compare the proposed model CAMM against four baseline models under two categories. Firstly, we compare our model with unimodal models, Bi-LSTM for text and DCNN for image classification. Secondly, we compare CAMM with the two popular multimodal fusion networks, MUTAN and BLOCK. Lastly, CAMM is also compared with other state-of-the-art multimodal models proposed in the recent disaster-related studies.

VI. IMPLEMENTATION DETAILS

In the following subsections, we discuss the implementation details of the unimodal and multimodal models for the binary classification of text and image tweets.

A. UNIMODAL CLASSIFICATION OF TEXT TWEETS

Fig. 4 shows the architecture for the unimodal classification of text tweets by the Bi-LSTM model.

1) PREPROCESSING OF TEXT

The tweets are preprocessed by removing the stop words and URLs embedded in the tweet text. After this, all the special ('@', '#', '$', '%', '&', '*', '!') and non_ASCII characters and their repeated occurrence are removed. Trailing and multiple white spaces in between are also deleted. Lastly, the text is converted into lowercase.

2) WORD EMBEDDINGS USING PRE-TRAINED GloVe

After the preprocessing step, the words in the tweet need to be represented as real-valued vectors for further processing. We have used pre-trained word embedding GloVe (Global Vectors for Word Representation) [56] to get the word embeddings. GloVe converts the words into vectors so that similar words have similar vector representations. To capture complete information from the word, we used the GloVe embedding of dimension 300.

3) CLASSIFICATION USING LSTM

Once the vector matrix of tweet words is obtained, we use the Bi-LSTM model to extract the features for classification. This study uses two LSTMs, one in the forward and one in the backward direction. LSTM model was first proposed by Hochreiter *et al.* [57] to handle the shortcomings of Recurrent Neural Networks, which could not handle the long-term dependencies. LSTMs are designed to remember the information for a longer time through a series of LSTM units. Each unit of the LSTM has a forget gate, input gate, output gate, and cell state. The forget gate consists of a sigmoid function that outputs a number between 0 and 1 depending on the previous and the current state. A '0' represents discard or forget, and a '1' represents keep or remember. The input gate also has a sigmoid function that decides which values to be updated, and the tanh function provides the new updated values resulting in the output for the next hidden state. These gates allow the model to keep only the critical information and forget the rest. We have used two layers of LSTM, which

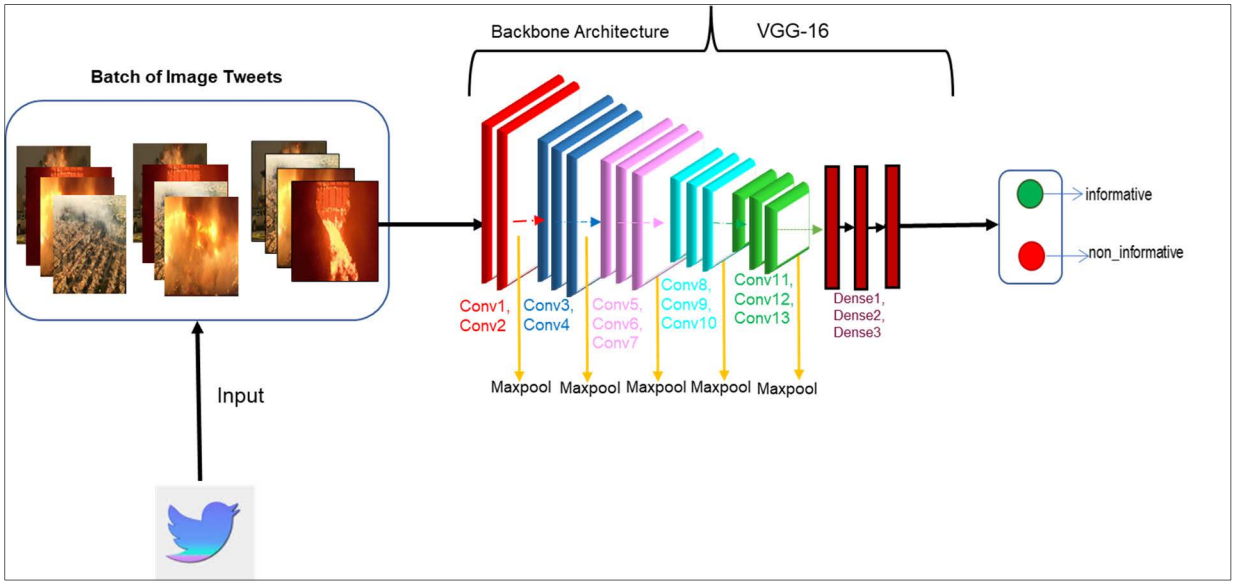

FIGURE 5. Unimodal classification of image tweets into two classes '*informative*' and '*non_informative*' using DCNN with VGG-16 as the backbone architecture.

are enough to capture the complex features from the tweet text.

B. UNIMODAL CLASSIFICATION OF IMAGE TWEETS

Fig. 5 shows the architecture of the unimodal classification of image tweets using DCNN with pre-trained VGG-16 [58] model as the backbone architecture.

1) PREPROCESSING OF IMAGES

We first resize the tweet images to (256, 256), and then a random patch of (224, 224) is cropped. After this, the pixels are scaled between 0 and 1 and then normalized with mean and standard deviation values of the ImageNet database [59].

2) CLASSIFICATION USING DEEP CONVOLUTIONAL NEURAL NETWORK (DCNN)

DCNN is used to classify tweet images using VGG-16 as the backbone architecture. VGG-16 is a DCNN designed to classify 14 million annotated images of the ImageNet dataset into 1000 classes. It consists of a stack of 13 convolutional layers with a small filter size of (3×3) divided into five blocks of 2, 2, 3, 3, 3 convolution layers where each block is followed by a maxpooling of (2×2) and a stride of 2 which reduces the size of the image to half after each block. After the five blocks of convolution layers there are three fully connected (FC) layers with 4096 parameters for the first two FC layers. The last FC has 1000 parameters which are equal to the number of classes for the ImageNet dataset.

Since we have two classes for the CrisisMMD dataset, we set the parameters of the last FC to two. In the end, a SoftMax function gives the probability of the output classes. After each hidden layer, the nonlinearity is introduced by adding an activation function ReLU.

C. MULTIMODAL CLASSIFICATION OF TEXT AND IMAGE TWEETS

We follow the same preprocessing technique for CAMM as mentioned in unimodal classification for text and image data. The output of the convolution layers of VGG-16 for an input image is of dimensions (7, 7, 512) which are passed through an additional 1×1 convolution and Tanh() activation to increase the number of channels to 1024. Each cell in the 7×7 matrix represents features for different spatial regions in the input image, which are subsequently reshaped to 49×1024 which represents F_I.

Bi-LSTM takes the GloVe embeddings as input for all the words in a tweet to generate features of dimension 1024 for each word represented as $F_T = (n, 1024)$ where n is the number of words in a tweet. The two feature vectors $F_I a$ nd F_T are then used to generate the key, query, and value vectors where W_K, W_Q and W_V are three separate fully connected layers with input and output size 1024. Next, the key and query vectors are used to generate the attention map M (refer to 1), which represents the relevance of each word in a tweet against different spatial regions of the image. Finally, this attention map is applied to the value vector to obtain a fused feature vector F_{CA} (refer to 2). The final fused vector is passed through a linear classifier consisting of three fully connected layers with a ReLU activation in between, and the output of the last layer is passed through a SoftMax function to get the probabilities assigned to each label.

VII. RESULTS AND DISCUSSION

To validate the performance of the proposed model CAMM, we conduct extensive experiments on the benchmark multimodal disaster dataset CrisisMMD and compare the results with baseline unimodal and multimodal methods and also

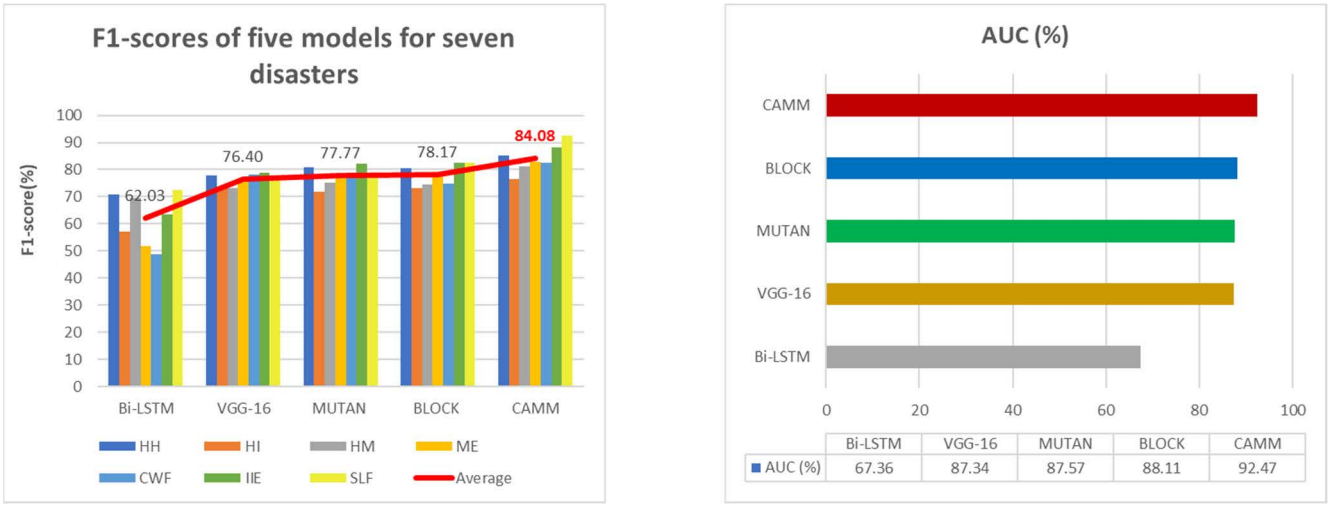

FIGURE 6. Chart (a) shows model wise average F1-score for five models. Chart (b) compares the Area Under the Curve (AUC), taken as the average of all the disasters for five models.

with recent state-of-the-art multimodal models. In this section, we present the results obtained after training the network under five setups: (1) Text-only, (2) Image-only, (3) MUTAN fusion, (4) BLOCK fusion, and (5) CAMM fusion (proposed). Under each setup, we train networks for seven disasters resulting in 35 test cases. Table 5 summarizes the Accuracy, Precision, Recall, F1-score, and AUC of the models trained under these five setups for seven disasters. Table 6 compares the results of CAMM with current state-of-the-art models. For comparison, we take the average of the F1-scores for all seven disasters for each model.

A. CAMM VS. BASELINE UNIMODAL AND MULTIMODAL MODELS

Firstly, we discuss the performance of the baseline unimodal and multimodal models trained in this study. The image-only DCNN architecture with VGG-16 as the backbone outperforms the text-only model Bi-LSTM. The average F1-score for the unimodal text-only model is 62.03% and for the image-only model is 76.40%. The performance of both the multimodal models MUTAN and BLOCK is better than the two unimodal models. For the MUTAN fusion model, we achieve an average F1-score of 77.77%, and for the BLOCK fusion model, it is 78.17%. BLOCK fusion method's performance is 0.4% better than MUTAN fusion. The proposed CAMM model outperforms all four models by achieving the average F1-score of 84.08% for all disasters, as shown in Fig. 6(a). Thus, we can see a clear progression from unimodal to multimodal classifiers, with the best F1-score achieved by CAMM. We also compare the five model's AUC metric taken as the average of all disasters. Fig. 6(b) confirms the outperformance of CAMM as it has a maximum value of 92.47% for AUC. The F1-scores of CAMM for individual disasters are Hurricane Harvey: 85.29%, Hurricane Irma:76.40%, Hurricane Maria: 81.01%, Mexico Earthquake: 82.93%, California Wildfires: 82.36%, Iraq-Iran Earthquake: 88.14%, and SriLanka Floods: 92.47%.

TABLE 2. Performance of CAMM for five backbone architectures.

Backbone Architecture	Accuracy	Precision	Recall	F1-Score
EfficientNet-B3	80.76	77.73	77.50	77.33
ResNet50	79.63	77.61	77.39	77.28
DenseNet201	79.19	80.22	71.63	75.49
VGG-16	**77.53**	**78.28**	**78.19**	**78.19**
VGG-19	77.37	77.43	70.62	73.67

TABLE 3. Performance of CAMM for different learning rates.

Learning Rate	Accuracy	Precision	Recall	F1-Score
1.00e-02	86.59	84.12	83.12	83.60
1.00e-03	**87.70**	**86.80**	**82.65**	**84.33**
1.00e-04	85.49	83.27	80.77	81.85
1.00e-05	77.66	35.33	50.00	41.40

TABLE 4. Performance of CAMM for different weight decay values.

Weight Decay	Accuracy	Precision	Recall	F1-Score
1.00e-03	70.66	35.33	50.00	41.40
1.00e-04	86.91	85.06	82.56	83.65
5.00e-04	**87.85**	**86.18**	**83.86**	**84.89**

B. CAMM VS. RECENT STATE-OF-THE-ART MULTIMODAL MODELS

We also compare the results of CAMM with recent state-of-the-art multimodal models in Table 6. The multimodal fusion

TABLE 5. Results of two Unimodal Models (1) Bi-LSTM for text tweets and (2) VGG-16 for image tweets and three Multimodal Models (1) MUTAN (2) BLOCK and the proposed model (3) CAMM on the data of seven disasters of CrisisMMD Dataset.

Disasters	Unimodal Models										Multimodal Models														
	Bi-LSTM(Text)					VGG-16(Images)					MUTAN					BLOCK					CAMM (Proposed)				
	Acc.	P	R	F1-score	AUC	Acc.	P	R	F1-score	AUC	Acc.	P	R	F1-score	AUC	Acc.	P	R	F1-score	AUC	Acc.	P	R	F1-score	AUC
Hurricane Harvey	77.37	72.84	69.84	70.93	77.94	81.94	78.54	77.24	77.77	87.32	84.78	82.45	79.80	80.91	89.06	84.55	82.26	79.24	80.49	89.34	88.33	87.23	83.88	85.29	91.84
Hurricane Irma	70.89	60.55	56.98	57.17	64.71	78.49	73.34	73.78	73.43	83.28	79.47	75.13	70.17	71.79	81.50	80.27	76.14	71.62	73.21	83.11	83.04	81.04	74.14	76.40	88.65
Hurricane Maria	70.90	70.14	69.26	69.51	76.07	75.00	75.23	72.78	73.24	84.12	76.61	76.79	74.60	75.11	83.97	76.13	76.34	74.02	74.53	85.27	82.24	83.41	80.24	81.01	89.69
Mexico Earthquake	63.56	52.87	52.21	51.83	52.62	81.33	78.56	75.58	76.69	89.82	82.00	78.92	77.52	78.14	89.21	82.89	80.02	78.57	79.21	88.66	86.23	85.33	81.59	82.93	92.03
California Wildfires	69.71	52.80	51.13	48.81	48.21	84.24	81.57	76.35	78.29	87.74	84.03	81.40	75.96	77.97	90.28	82.37	79.64	72.65	74.90	88.44	87.16	84.89	80.62	82.36	92.19
Iraq Iran Earthquake	78.00	63.12	63.82	63.43	73.07	87.00	78.01	80.15	78.90	91.22	88.50	80.01	85.40	82.22	92.41	89.00	80.80	84.62	82.48	92.02	92.00	84.69	94.04	88.14	95.93
Sri Lanka Floods	79.19	73.75	71.71	72.57	78.91	82.08	77.89	75.69	76.47	87.92	81.50	77.02	80.63	78.24	86.60	85.55	81.55	83.75	82.39	89.92	94.22	94.36	91.03	92.47	96.99
Average	72.80	63.72	62.13	62.03	67.36	81.44	77.59	75.94	76.40	87.34	82.41	78.81	77.72	77.77	87.57	82.96	79.53	77.78	78.17	88.11	**87.60**	**85.85**	**83.65**	**84.08**	**92.47**

model with logistic regression proposed by Gautam *et al.* [25] reported only their accuracy results. Their model achieves accuracy in the range of 74.14% to 80.20% for the seven disasters with an average accuracy of 77.33% compared to CAMM, which has an average accuracy of 87.60%, thus confirming the advantage of using an attention mechanism. Ofli's early fusion model [26] gives an F1-score of 84.2% for the informativeness task, which is close to CAMM. Although they worked on the CrisisMMD dataset, one significant difference in their training is that they merged the data of all seven disasters to get a much bigger dataset which directly impacted the performance of their deep learning model. In a similar study, models trained by Caragea *et al.* [27] for late fusion with majority voting and early fusion give the average F1-score of 75.63% and 78.13%, respectively. Two other studies that used the early and late fusion with majority voting show the F1-score of 74.46% by Madichetty *et al.* [28] and 82% by Kumar *et al.* [29] as compared to CAMM's F1-score of 84.08%.

The above discussion confirms the following:

1) The performance of multimodal classification models is better than unimodal text-only and image-only models for all seven disasters of the CrisisMMD dataset.
2) Amongst the three multimodal techniques, the proposed cross-attention-based model CAMM outperforms the MUTAN and BLOCK models.
3) CAMM also outperforms current state-of-the-art multimodal models on the benchmark CrisisMMD dataset, confirming the advantage of applying cross-attention fusion for text and image modalities. The proposed method is designed to select the prominent features from the two modalities which are more relevant for the task resulting in a better classifier.

We have used the following abbreviations for naming the disasters:

HH: Hurricane Harvey, **HI**: Hurricane Irma,
HM: Hurricane Maria, **ME**: Mexico Earthquake,
CWF: California Wildfires, **IIE**: Iraq-Iran Earthquake, **SLF**: Sri Lanka Floods

C. ABLATION STUDY

For the proposed CAMM architecture, the hyperparameters are fine-tuned using grid search. The results of experiments performed for selecting the backbone architecture are shown in Table 2. We trained the network for Hurricane Harvey image dataset with EfficientNet-B3, ResNet50, DenseNet201, VGG-16 and VGG-19 as backbone architectures. The results confirm that VGG-16 achieves the highest F1-score of 78.19% and hence the best choice for all the experiments performed in this study.

The results of fine-tuning the learning rate on Hurricane Harvey image dataset are listed in Table 3. Out of [1.00e-02, 1.00e-03, 1.00e-04, 1.00e-05] the best F1-score of 84.33% is achieved with 1.00e-03.

Similarly, we trained the network on the Hurricane Harvey image dataset for the values [1.00e-03, 1.00e-04, 5.00e-04] for the selection of weight decay hyperparameter and chose 5.00e-04 for all our experiments in this study. The results are listed in Table 4.

TABLE 6. Comparison of CAMM with recent state-of-the-art multimodal models on CrisisMMD Dataset. The symbols used for various disasters are: HH: Hurricane Harvey, HI: Hurricane Irma, HM: Hurricane Maria, ME: Mexico Earthquake, CWF: California Wildfires, IIE: Iraq-Iran Earthquake, and SLF: Sri Lanka Floods.

Ref.	Approach	Dataset & Classes	Disaster	Accuracy	Precision	Recall	F1-score
[25]	Logistic Regression based decision policy for multimodal analysis with LSTM for text and ResNet50 for images. *(Late or decision Fusion)*	CrisisMMD [33] informative / non-informative	HH	79.20	-	-	-
			HI	80.20			
			HM	79.40			
			ME	77.90			
			CWF	75.30			
			IIE	75.20			
			SLF	74.14			
			AVERAGE	77.33			
[26]	Two parallel networks are trained, for text CNN with 5 hidden layers and for images, VGG-16 is used. The two feature vectors are combined and passed to a dense layer and then a SoftMax layer. *(Early Fusion)*	CrisisMMD [33] (i) informative / non-informative (ii) Humanitarian Categories	Combined datasets of Seven Disasters: HH, HI, HM, ME, CWF, IIE, SLF	84.4	84.1	84.0	84.2
				78.4	78.5	78.0	78.3
[27]	The output of two independent networks RNN(GRU) for text and CNN for images are combined in two ways: (1) Average of the probabilities of the predictions of two networks *(Late Fusion)* (2) Concatenate the outputs of the last layers of two networks and feed to the dense layer to make the final prediction *(Early Fusion)*	CrisisMMD [33] informative / non-informative	HH	79.32,79.78			78.17,79.20
			HI	74.98,79.91	-	-	72.26,79.07
			HM	74.23,76.67			73.89,76.59
			ME	76.27,73.62	-	-	73.27,72.92
			CWF	78.50,75.45			76.04,73.20
			IIE	76.15,75.06			71.89,74.13
			SLF	84.60,91.59			83.90,91.85
			AVERAGE	77.72,78.87			75.63,78.13
[28]	Text tweets are passed to CNN and image tweets are trained using VGG-16. The probability vector of these two networks is combined using additive fusion. The final class label is determined by majority voting. *(Late Fusion + majority voting)*	CrisisMMD [33] informative / non-informative	HH	77.70	78.00	78.00	77.60
			HI	73.82	74.00	74.00	73.55
			HM	72.96	73.00	73.00	72.84
			ME	74.29	74.50	74.00	74.25
			CWF	65.00	65.50	64.50	64.00
			IIE	68.18	71.00	68.00	67.00
			SLF	91.78	92.00	91.50	92.00
			AVERAGE	74.81	75.42	74.71	74.46
[29]	For text tweets LSTM and for image tweets VGG-16 is used. For the multimodal training the features of text and image modalities are concatenated. Final label is assigned based on the majority voting of text, image and multimodal labels. *(Early Fusion + majority voting)*	CrisisMMD [33] informative / non-informative	HH	-	84.00	84.00	84.00
			HI	-	82.00	82.00	82.00
			HM	-	84.00	84.00	84.00
			ME	-	75.00	74.00	74.00
			CWF	-	75.00	74.00	74.00
			IIE	-	84.00	83.00	83.00
			SLF	-	93.00	93.00	93.00
			AVERAGE		82.42	82.00	82.00
Baseline Model	MUTAN fusion (multimodal)	CrisisMMD [33] informative / non-informative	HH	84.78	82.45	79.80	80.91
			HI	79.47	75.13	70.17	71.79
			HM	76.61	76.79	74.60	75.11
			ME	82.00	78.92	77.52	78.14
			CWF	84.03	81.40	75.96	77.97
			IIE	88.50	80.01	85.40	82.22
			SLF	81.50	77.02	80.63	78.24
			AVERAGE	82.41	78.81	77.72	77.77
Baseline Model	BLOCK fusion (multimodal)	CrisisMMD [33] informative / non-informative	HH	84.55	82.26	79.24	80.49
			HI	80.27	76.14	71.62	73.21
			HM	76.13	76.34	74.01	74.53
			ME	82.89	80.02	78.57	79.21
			CWF	82.37	79.64	72.65	74.90
			IIE	89.00	80.80	84.62	82.48
			SLF	85.55	81.55	83.75	82.39
			AVERAGE	82.96	79.53	77.78	78.17
Proposed Model	A novel Cross-Attention Multi Modal (CAMM) fusion model *(Cross-Attention Fusion)*	CrisisMMD [33] informative / non-informative	HH	88.33	87.23	83.88	85.29
			HI	83.04	81.04	74.14	76.40
			HM	82.24	83.41	80.24	81.01
			ME	86.23	85.33	81.59	82.93
			CWF	87.16	84.89	80.62	82.36
			IIE	92.00	84.69	94.04	88.14
			SLF	94.22	94.36	91.03	92.47
			AVERAGE	**87.60**	**85.85**	**83.65**	**84.08**

VIII. CONCLUSION

In this study, we proposed a novel Cross-Attention Multi-Modal (CAMM) framework for the binary classification of multimodal data of seven disasters collected from the Twitter microblogging platform. A series of experiments performed under five different setups confirm that the proposed multimodal classifier that uses the complementary features of textual and visual modalities performs better than the unimodal text-only and image-only approaches. Also, the cross-attention fusion mechanism performs better than the previously proposed methods based on early and late fusion techniques. As a result, CAMM achieved an average F1-score of 84.08%, which is 6.31% better than the F1-score of the MUTAN fusion method and 5.91% better than the BLOCK fusion method trained in this study. CAMM also outperformed recent state-of-the-art models for the benchmark multimodal disaster dataset. In the future, we would like to extend the concept of cross-attention to other modalities like audio and video. We would also explore how to use the attention mechanism for transfer learning and domain adaptation in scenarios where labeled data is unavailable, especially at the onset of a new disaster.

REFERENCES

[1] C. M. Vera-burgos and D. R. Grif, "Using Twitter for crisis communications in a natural disaster?: Hurricane Harvey," *Heliyonvol*, vol. 6, p. 9, Aug. 2020, doi: 10.1016/j.heliyon.2020.e04804.

[2] J. Yang, M. Yu, H. Qin, M. Lu, and C. Yang, "A Twitter data credibility framework—Hurricane Harvey as a use case," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 3, p. 111, Feb. 2019, doi: 10.3390/ijgi8030111.

[3] J. Tang, S. Yang, and W. Wang, "Social media-based disaster research: Development, trends, and obstacles," *Int. J. Disaster Risk Reduction*, vol. 55, Mar. 2021, Art. no. 102095, doi: 10.1016/j.ijdrr.2021.102095.

[4] A. V Mavrodieva and R. Shaw, *Social Media in Disaster Management*. Singapore: Springer, 2021, doi: 10.1007/978-981-16-0285-6.

[5] J. Phengsuwan, T. Shah, N. B. Thekkummal, Z. Wen, R. Sun, D. Pullarkatt, H. Thirugnanam, M. V. Ramesh, G. Morgan, P. James, and R. Ranjan, "Use of social media data in disaster management: A survey," *Futur. Internet*, vol. 13, no. 2, pp. 1–24, 2021, doi: 10.3390/fi13020046.

[6] C. Reuter and M.-A. Kaufhold, "Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics," *J. Contingencies Crisis Manage.*, vol. 26, no. 1, pp. 41–57, Mar. 2018, doi: 10.1111/1468-5973.12196.

[7] J. K. Joseph, K. A. Dev, A. P. Pradeepkumar, and M. Mohan, "Big data analytics and social media in disaster management," in *Integrating Disaster Science and Management: Global Case Studies in Mitigation and Recovery*. 2018, doi: 10.1016/B978-0-12-812056-9.00016-6.

[8] A. Khattar and S. M. K. Quadri, "Emerging role of artificial intelligence for disaster management based on microblogged communication," in *Proc. Int. Conf. Innov. Comput. Commun. (ICICC)*, Mar. 2020, doi: 10.2139/ssrn.3562973.

[9] H. Purohit, C. Castillo, M. Imran, and R. Pandey, "Social-EOC: Serviceability model to rank social media requests for emergency operation centers," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 119–126, doi: 10.1109/ASONAM.2018.8508709.

[10] S. Ghosh and M. S. Desarkar, "Class specific TF-IDF boosting for short-text classification," in *Proc. Companion Web Conf. Web Conf. (WWW)*, 2018, pp. 1629–1637, doi: 10.1145/3184558.3191621.

[11] M. Basu, A. Shandilya, P. Khosla, K. Ghosh, and S. Ghosh, "Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 3, pp. 604–618, Jun. 2019, doi: 10.1109/TCSS.2019.2914179.

[12] K. Rudra, S. Ghosh, N. Ganguly, P. Goyal, and S. Ghosh, "Extracting situational information from microblogs during disaster events," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 583–592, doi: 10.1145/2806416.2806485.

[13] F. Alam, S. Joty, and M. Imran, "Domain adaptation with adversarial training and graph embeddings," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 1077–1087, doi: 10.18653/v1/P18-1099.

[14] F. Alam, F. Ofli, and M. Imran, "Processing social media images by combining human and machine computing during crises," *Int. J. Hum.–Comput. Interact.*, vol. 34, no. 4, pp. 311–327, Apr. 2018, doi: 10.1080/10447318.2018.1427831.

[15] X. Li, C. Caragea, D. Caragea, M. Imran, and F. Ofli, "Identifying disaster damage images using a domain adaptation approach," in *Proc. Int. ISCRAM Conf.*, May 2019, pp. 633–645. [Online]. Available: https://par.nsf.gov/servlets/purl/10204518.

[16] A. Khattar and S. M. K. Quadri, "Deep domain adaptation approach for classification of disaster images," in *Intelligent Data Communication Technologies and Internet of Things* (Lecture Notes on Data Engineering and Communications Technologies), vol. 57. Singapore: Springer, 2021, pp. 245–259, doi: 10.1007/978-981-15-9509-7_21.

[17] S. Pouyanfar, Y. Tao, H. Tian, S.-C. Chen, and M.-L. Shyu, "Multimodal deep learning based on multiple correspondence analysis for disaster management," *World Wide Web*, vol. 22, no. 5, pp. 1893–1911, Sep. 2019, doi: 10.1007/s11280-018-0636-4.

[18] H. Singh, A. Nasery, D. Mehta, A. Agarwal, J. Lamba, and B. V. Srinivasan, "MIMOQA: Multimodal input multimodal output question answering," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 5317–5332.

[19] S. Zhang, M. Chen, J. Chen, F. Zou, Y.-F. Li, and P. Lu, "Multimodal feature-wise co-attention method for visual question answering," *Inf. Fusion*, vol. 73, pp. 1–10, Sep. 2021, doi: 10.1016/j.inffus.2021.02.022.

[20] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *J. Appl. Sci. Technol. Trends*, vol. 2, no. 2, pp. 52–58, Apr. 2021, doi: 10.38094/jastt20291.

[21] A. Kumar and J. Vepa, "Gated mechanism for attention based multi modal sentiment analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4477–4481, doi: 10.1109/ICASSP40776.2020.9053012.

[22] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3025–3035. [Online]. Available: https://aclanthology.org/2020.acl-main.273

[23] G. Cai, J. Zhang, X. Jiang, Y. Gong, L. He, F. Yu, P. Peng, X. Guo, F. Huang, and X. Sun, "Ask & confirm: Active detail enriching for cross-modal retrieval with partial query," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1835–1844.

[24] H. Zhang, "Voice keyword retrieval method using attention mechanism and multimodal information fusion," *Sci. Program.*, vol. 2021, pp. 1–11, Jan. 2021, doi: 10.1155/2021/6662841.

[25] A. K. Gautam, L. Misra, A. Kumar, K. Misra, S. Aggarwal, and R. R. Shah, "Multimodal analysis of disaster tweets," in *Proc. IEEE 5th Int. Conf. Multimedia Big Data (BigMM)*, Sep. 2019, pp. 94–103, doi: 10.1109/BigMM.2019.00-38.

[26] F. Ofli and M. Imran, "Analysis of social media data using multimodal deep learning for disaster response," in *Proc. 17th ISCRAM Conf.*, May 2020, pp. 1–10. [Online]. Available: http://idl.iscram.org/files/ferdaofli/2020/2272_FerdaOfli_etal2020.pdf

[27] X. Li and D. Caragea, "Improving disaster-related Tweet classification with a multimodal approach," in *Proc. 17th ISCRAM Conf.*, May 2020, pp. 893–902.

[28] S. Madichetty and S. M, "Classifying informative and non-informative tweets from the Twitter by adapting image features during disaster," *Multimedia Tools Appl.*, vol. 79, nos. 39–40, pp. 28901–28923, Oct. 2020, doi: 10.1007/s11042-020-09343-1.

[29] A. Kumar, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, *A Deep Multi-Modal Neural Network for Informative Twitter Content Classification During Emergencies*, no. 0123456789. USA: Springer, 2020, doi: 10.1007/s10479-020-03514-x.

[30] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5999–6009.

[32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–15.

[33] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter datasets from natural disasters," in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, 2018, pp. 1–9. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17816/17038

[34] H. Ben-younes, R. Cadene, M. Cord, and N. Thome, "MUTAN: Multimodal tucker fusion for visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2631–2639, doi: 10.1109/ICCV.2017.285.

[35] H. Ben-younes, R. Cadene, N. Thome, and M. Cord, "BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, 2019, pp. 8102–8109.

[36] S. Goswami, S. Chakraborty, S. Ghosh, A. Chakrabarti, and B. Chakraborty, "A review on application of data mining techniques to combat natural disasters," *Ain Shams Eng. J.*, vol. 9, no. 3, pp. 365–378, Sep. 2018, doi: 10.1016/j.asej.2016.01.012.

[37] H. Zade, K. Shah, V. Rangarajan, P. Kshirsagar, M. Imran, and K. Starbird, "From situational awareness to actionability: Towards improving the utility of social media data for crisis response," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, pp. 1–18, Nov. 2018, doi: 10.1145/3274464.

[38] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Extracting information nuggets from disaster-related messages in social media," in *Proc. 10th Int. Conf. Inf. Syst. Cris. Response Manag. (ISCRAM)*, May 2013, pp. 1–10.

[39] M. Basu, A. Shandilya, K. Ghosh, and S. Ghosh, "Automatic matching of resource needs and availabilities in microblogs for post-disaster relief," in *Proc. Companion Web Conf. WWW*, Jan. 2019, pp. 25–26, doi: 10.1145/3184558.3186911.

[40] S. Madichetty and S. M, "A stacked convolutional neural network for detecting the resource tweets during a disaster," *Multimedia Tools Appl.*, vol. 80, no. 3, pp. 3927–3949, Jan. 2021, doi: 10.1007/s11042-020-09873-8.

[41] R. Suwaileh, M. Imran, T. Elsayed, and H. Sajjad, "Are we ready for this disaster? Towards location mention recognition from crisis tweets," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6252–6263, doi: 10.18653/v1/2020.coling-main.550.

[42] F. K. Sufi and I. Khalil, "Automated disaster monitoring from social media posts using AI-based location intelligence and sentiment analysis," *IEEE Trans. Computat. Social Syst.*, early access, Mar. 18, 2022, doi: 10.1109/TCSS.2022.3157142.

[43] H. M. Zahera, R. Jalota, M. A. Sherif, and A.-C.-N. Ngomo, "I-AID: Identifying actionable information from disaster-related tweets," *IEEE Access*, vol. 9, pp. 118861–118870, 2021, doi: 10.1109/ACCESS.2021.3107812.

[44] R. McCreadie, C. Buntain, and I. Soboroff, "TREC incident streams: Finding actionable information on social media," in *Proc. Int. ISCRAM Conf.*, May 2019, pp. 691–705.

[45] I. Triguero, S. García, and F. Herrera, "Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study," *Knowl. Inf. Syst.*, vol. 42, no. 2, pp. 245–284, 2015, doi: 10.1007/s10115-013-0706-y.

[46] N. Pandey and S. Natarajan, "How social media can contribute during disaster events? Case study of Chennai floods 2015," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2016, pp. 1352–1356, doi: 10.1109/ICACCI.2016.7732236.

[47] I. Triguero, S. Garcia, and F. Herrera, "SEG-SSC: A framework based on synthetic examples generation for self-labeled semi-supervised classification," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 622–634, Apr. 2015, doi: 10.1109/TCYB.2014.2332003.

[48] F. Alam, S. Joty, and M. Imran, "Graph based semi-supervised learning with convolution neural networks to classify crisis related Tweets," in *Proc. 12th Int. AAAI Conf. Web Soc. Media (ICWSM)*, May 2018, pp. 556–559.

[49] H. Li, N. Guevara, N. Herndon, D. Caragea, K. Neppalli, C. Caragea, A. C. Squicciarini, and A. H. Tapia, "Twitter mining for disaster response: A domain adaptation approach," in *Proc. 12th Int. Conf. Inf. Syst. Cris. Response Manag. (ISCRAM)*, Jan. 2015, pp. 1–7. [Online]. Available: https://people.cs.ksu.edu/ ccaragea/papers/iscram15.pdf

[50] H. Li, D. Caragea, C. Caragea, and N. Herndon, "Disaster response aided by Tweet classification with a domain adaptation approach," *J. Contingencies Crisis Manage.*, vol. 26, no. 1, pp. 16–27, Mar. 2018, doi: 10.1111/1468-5973.12194.

[51] D. T. Nguyen, K. A. Al Mannai, S. Joty, H. Sajjad, M. Imran, and P. Mitra, "Rapid classification of crisis-related data on social networks using convolutional neural networks," in *Proc. 11th Int. Conf. Web Soc. Media (ICWSM)*, Aug. 2016, pp. 632–635.

[52] S. Ahadzadeh and M. R. Malek, "Earthquake damage assessment in three spatial scale using naive Bayes, SVM, and deep learning algorithms," *Appl. Sci.*, vol. 11, no. 20, p. 9737, Oct. 2021, doi: 10.3390/app11209737.

[53] A. Khattar and S. M. K. Ouadri, "A semi-supervised domain adaptation approach for diagnosing SARS-CoV-2 and its variants of concern (VOC)," in *Proc. 9th Int. Conf. Rel., INFOCOM Technol. Optim. (Trends Future Directions) (ICRITO)*, Sep. 2021, pp. 1–9, doi: 10.1109/ICRITO51393.2021.9596381.

[54] B. W. Robertson, M. Johnson, D. Murthy, W. R. Smith, and K. K. Stephens, "Using a combination of human insights and 'deep learning' for real-time disaster communication," *Prog. Disaster Sci.*, vol. 2, Jul. 2019, Art. no. 100030, doi: 10.1016/j.pdisas.2019.100030.

[55] W. Nie, Q. Liang, Y. Wang, X. Wei, and Y. Su, "MMFN," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 4, pp. 1–22, Jan. 2021, doi: 10.1145/3410439.

[56] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Sep. 2014, pp. 1–14.

[59] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database," *J. Vis.*, vol. 9, no. 8, p. 1037, 2010, doi: 10.1167/9.8.1037.

ANURADHA KHATTAR received the M.Sc. degree in mathematics from the University of Delhi, India, and the Master of Computer Applications degree from the IGNOU, India. She is currently pursuing the Ph.D. degree with Jamia Millia Islamia, New Delhi, India. She is also an Associate Professor with the Department of Computer Science, Miranda House, University of Delhi. Her research interests include deep learning and computer vision applied in the field of disaster management.

S. M. K. QUADRI received the M.Tech. degree in computer applications from the Indian School of Mines (ISM), Dhanbad, India, in 1998, and the Ph.D. degree in computer science from the University of Kashmir, India, in 2008. He is currently a Professor with the Department of Computer Science, Jamia Millia Islamia, India. He has more than 30 years of teaching and research experience. He has more than 150 publications in international and national peer-reviewed journals and conferences to his credit.