IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 June 2023, accepted 7 July 2023, date of publication 12 July 2023, date of current version 19 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3294474

RESEARCH ARTICLE

Categorizing Crises From Social Media Feeds via Multimodal Channel Attention

MARIHAM REZK[1,2], **NOURELDIN ELMADANY**[1,2], **RADWA K. HAMAD**[1], **AND EHAB F. BADRAN**[1], **(Senior Member, IEEE)**

[1]Department of Electronics and Communications Engineering, Arab Academy for Science, Technology and Maritime Transport, Alexandria 21937, Egypt
[2]Intelligent Systems Laboratory, Arab Academy for Science, Technology and Maritime Transport, Alexandria 21937, Egypt

Corresponding author: Noureldin Elmadany (nourelmadany@aast.edu)

ABSTRACT In the era of advanced computer vision and natural language processing, the use of social media as a source of information has become even more valuable in directing aid and rescuing victims. Consequently, millions of texts and images can be processed in real-time, allowing emergency responders to efficiently assess evolving crises and appropriately allocate resources. The majority of the previous detection studies are text-only or image-only based, overlooking the potential benefits of integrating both modalities. In this paper, we propose Multimodal Channel Attention (MCA) block, which employs an adaptive attention mechanism, learning to assign varying importance to each modality. We then propose a novel Deep Multimodal Crisis Categorization (DMCC) framework, which employs a two-level fusion strategy for better integration of textual and visual information. The DMCC framework consists of feature-level fusion, which is accomplished through the MCA block, and score-level fusion, whereby the decisions made by the individual modalities are integrated with those of the MCA model. Extensive experiments on publicly available datasets demonstrate the effectiveness of the proposed framework. Through a comprehensive evaluation, it was found that the proposed framework achieves a performance enhancement compared to unimodal methods. Furthermore, it outperforms the current state-of-the-art methods on crisis-related categorization tasks. The code is available at https://github.com/MarihamR/Categorizing-Crises-from-Social-Media-Feeds-Via-Multimodal-Channel-Attention.

INDEX TERMS Multimodal deep learning, social media, natural disasters, crisis response, attention, fusion.

I. INTRODUCTION

Billions of posts are constantly shared every second on social media platforms, which encompass a broad range of significant events. In times of crisis, utilizing social media platforms can provide valuable and actionable information more efficiently than traditional emergency communication channels [1]. Crisis management utilizing social media data encompasses a plethora of endeavors, seeking to swiftly pinpoint and address emergency scenarios. One crucial aspect is crisis detection and identification, which encloses the automatic identification of posts suggesting a transpiring crisis [2], [3], [4], such as: natural disasters [5], [6], terrorist attacks [7], and public health emergencies [8], [9], [10]. A further task is sentiment analysis, which identifies and analyzes the emotional tone of crisis-related posts or tweets. This provides a deeper understanding of public reactions to the crisis [11], [12]. Additionally, crisis mapping task is also of paramount importance because it creates visual representations of crisis-related information, such as the location of affected areas and the spread of a disaster [13], [14]. Finally, the use of early warnings and alerts during disasters gained significant momentum recently. Because it allows individuals to disseminate information quickly and widely about the location and magnitude of a crisis, as well as providing guidance on evacuation routes and emergency shelter locations [15].

The field of crisis categorization and damage assessment through social media analysis is lacking in studies that employ multimodal methods. Such methods involve the integration of multiple forms of data, including text, images and

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

videos, to gain a more thorough understanding of a situation. Despite the abundance of information readily shared on social media platforms during crises, the majority of existing studies on damage assessment have traditionally focused on either image-based or text-based analysis. Thus, overlooking the complementary information that can be obtained from multiple modalities. Single modality methods in damage assessment suffer from several limitations, such as lack of context, ambiguity, bias, poor quality data, and limited information. Practically, each modality captures a certain kind of information that is likely to be complementary. For example, images provide on-site information from the eyewitness perspective which is hard to be described in words [16]. Therefore, integrating the information from multiple modalities via fusion is expected to improve the crisis categorization performance. Simple fusion techniques were adopted [17], [18], [19], [20]. However, they do not capture relations among the modalities.

To circumvent the above limitations, we first propose Multimodal Channel Attention (MCA) block for learning a common representation which discovers the interdependencies among the modalities. Finally, we introduce Deep Multimodal Crisis Categorization (DMCC) framework for integrating image and text data. The key contributions of the presented work are:

- The incorporation of attention mechanism in the context of deep multimodal learning is explored aiming to focus on the most pertinent information from each modality.
- MCA block is proposed as a feature level fusion technique, MCA block learns distinct weights for each modality, providing an effective approach for fusion.
- A DMCC framework comprising a two-level fusion strategy, intermediate fusion (MCA) and late fusion (score-level fusion), is introduced.
- Extensive experiments are conducted on two benchmark datasets: CrisisMMD [21] and DMD [17].
- The proposed framework shows superiority over the unimodal methods. Additionally, the evaluations demonstrate that DMCC framework outperforms the current state-of-the-art model by around 4 %, 5 %, and 1 % on CrisisMMD dataset task 1, CrisisMMD dataset task 2, and DMD dataset, respectively.
- The discriminatory capabilities of the proposed DMCC framework are assessed by conducting both quantitative and qualitative analysis.

The remaining part of the paper is organized as follows: In Section II, we review the most recent and relevant literature. Section III presents the details of the proposed MCA block and DMCC framework. Extensive experimental results are reported in Section IV. Finally, we conclude the paper in Section V.

II. RELATED WORK

Research on social media crisis identification spans nearly ten years proposing many methods that use different modalities. From modality type point of view, social media crisis identification methods can be categorized into two main groups: 1) Unimodal methods and 2) Multimodal methods.

A. UNIMODAL METHODS

There has been an extensive research conducted on crisis-related tasks from textual data. Some researchers adopted machine learning, graphical, and non-parametric models. In [22], Sakaki et al. presented a earthquake reporting system from tweets using Support Vector Machine (SVM). Imran et al. [23] utilized Naïve Bayesian (NB) classifier in identifying valuable information from disaster related tweets. In a subsequent work, the authors proposed a framework for detecting informative tweets using Conditional Random Field (CRF) [24]. Singh et al. [25] presented a markov model to predict the location victims. Shekhar and Setty [12] presented system to extract information about the disaster nature, emotions of affected people and relief efforts using K- Nearest Neighbour (KNN).

After the revolutionary impact of deep learning on computer vision tasks, researchers investigated using deep learning models, including CNN, Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM). For the task of identifying informative tweets during a disaster, Caragea et al. [26] used Bag of Words (BoW) and Text-CNN [27]. Similarly, Nguyen et al. [28] explored various word embedding techniques and utilized Text-CNN as for classifier. While Sreenivasulu et al. [29] combined Text-CNN with Feed Forward Neural Network (FFNN) to further improve the performance. In [30], Burel et al. also employed Text-CNN in crisis situations type detection. Alharbi and Lee [31] extracted crisis-related messages from Arabic Twitter data using different deep learning techniques, including Text-CNN, LSTM, and BiLSTM.

The majority of studies for assessing damage and disasters from images applied CNN. For example, Nguyen et al. [32] and Alam et al. [33] classified the damage severity using CNNs, While Li et al. [34] formulated the problem as regression. They trained a CNN to quantify the degree of the damage as a score. In [35], Kumar et al. also predicted the damage severity from twitter images. They used various CNN architectures for feature extraction, then applied different machine learning techniques, such as Support Vector Machines (SVM), Naïve Bayesian (NB), K- Nearest Neighbour (KNN), and Random Forest (RF), for the classification. Alam et al. in [36] and [37] addressed various crisis-related tasks using different CNN architectures in different tasks, such as disaster type detection, and informativeness classification, humanitarian categorization and damage severity assessment.

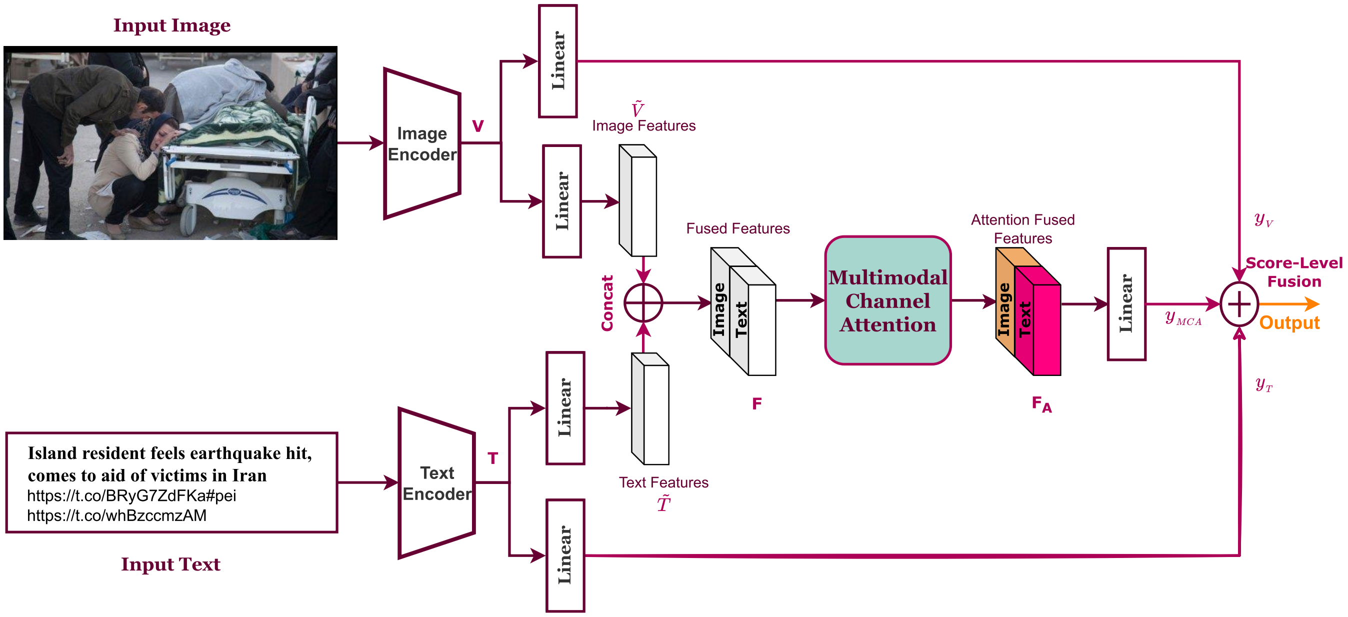


FIGURE 1. The proposed DMCC framework using MCA block considers two modalities, specifically text and image.

B. MULTIMODAL METHODS

The fusion of abundant and complementary information provides a common discriminative representation. There are currently a limited number of multimodal learning frameworks, developed specifically for crises related applications. Rizk et al. [38] proposed a two stage multimodal framework merging visual and semantic features to analyze twitter data during crisis. They used computationally inexpensive visual representations, including Gray-Level Co-Occurrence Matrix (GLCM) [39] and Gabor filters [40]. For text, they adopted BoW as a text descriptor. Mouzannar et al. [17] identified environmental and human damages from social media posts. They combined multiple unimodal CNNs that independently extract visual and textual feature representation. In [18], Hossain et al. demonstrated a multimodal damage detection system that extracted visual features from a pretrained ResNet50 [41] and textual features from a bidirectional long-term memory (BiLSTM) network with attention mechanism. They then concatenated visual and textual feature representation. Ofli et al. [20] adapted CNNs to learn a common modality-agnostic shared representations. Abavisani et al. [19] also presented a multimodal framework based on a cross-attention model. They used pretrained DenseNet [42] and BERT [43] to extract visual and textual embeddings, respectively. In [44], Gautam et al. presented simple decision fusion between visual and textual networks. In [45], Kumar et al. presented an end-to-end informativeness detector in tweets which concatenates image and text embedding.

III. PROPOSED DMCC FRAMEWORK

The proposed DMCC framework shown in Fig. 1, consists of four main modules. The first two modules are image and text encoders which are responsible for visual and textual feature representations, respectively. At the core of the framework is the proposed MCA block which fuses visual and textual modalities. Finally, the score fusion combines the decision of the previous three components.

A. DMCC IMAGE ENCODER

Images are important in crisis identification because they provide detailed on-site information [34]. In this study, the visual features are extracted using Convolutional Neural Networks (CNNs). We employ transfer learning technique to overcome the limited size of the available datasets [46]. In transfer learning, a pretrained CNN on a large dataset (i.e ImageNet dataset [47]) is fine-tuned on a smaller dataset. One of the efficient yet effective CNN architectures is EfficientNet [48] because it achieves high accuracy with few parameters. The main breakthrough of EfficientNet is based on highly effective compound scaling of width, depth and resolution. For the image encoder, we choose EfficientNetV2 [49], which enhances both training speed and parameter efficiency. The visual features are extracted from the final fully connected layer of EfficientNetV2 as follows:

$$V = EfficientNetV2(I), \tag{1}$$

where I is the input image, $V \in \mathbb{R}^{1000}$ is the output of the image encoder. The output is then split into two pathways. The first of which is a fully connected layer, which outputs the final classification scores for each class $\boldsymbol{y_V}$. The second pathway is for the feature-level fusion, which is a linear layer that acts as a visual features projection into a fixed dimension D. The linear layer consists of a fully connected layer, batch normalization and then an activation function,

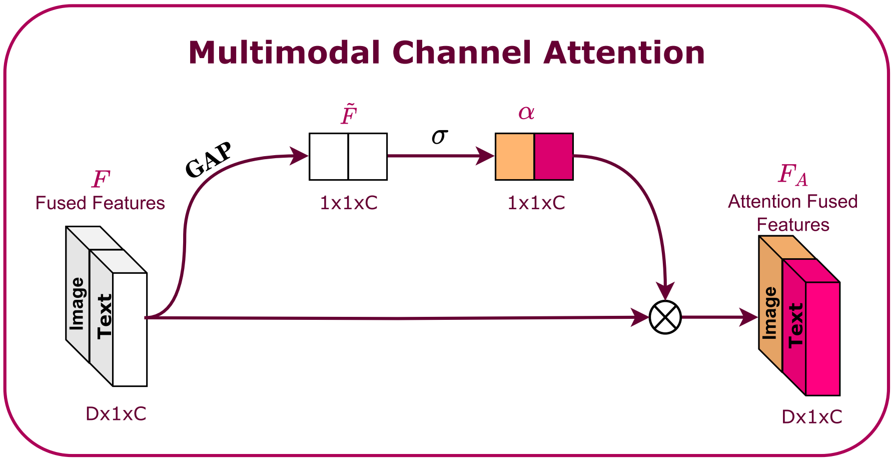


FIGURE 2. The proposed MCA block.

such as ReLU so that:

$$\tilde{V} = \delta(B_n(W_V' V)), \tag{2}$$

where $\tilde{V} \in \mathbb{R}^D$ is the visual features projection, W_V is the weight matrix of the linear layer, B_n is a batch normalization, and δ is a ReLU activation function. In section IV-B1, we conduct a comparative study among different image encoders candidates.

B. DMCC TEXT ENCODER

The choice of text encoder is crucial in the framework's performance. Transformers [50] are currently the state-of-the-art in natural language processing. One of the most widely used text encoders is BERT [43], a Google-developed bidirectional transformer with 110 million parameters. BERT is pretrained on a vast amount of text data. Here, BERT is fine-tuned on crisis-related tweets, a relatively small dataset, to improve the classification performance compared to random weight initialization. This can be attributed to the fact that the main textual patterns are already learnt during pretraining phase. In the proposed framework, BERT is used to extract the embeddings from each tweet sentence S as follows:

$$T = BERT(S), \tag{3}$$

where S is the input tweet text, $T \in \mathbb{R}^{768}$ is the output embedding of the text encoder. The output embedding of text encoder is divided into two pathways, conforming with the output of the image encoder. The first pathway yields the final classification scores for each class y_T through a fully connected layer. The second pathway is for the feature-level fusion. It consists of a linear layer acting as a textual features projection into a fixed dimension D as follows:

$$\tilde{T} = \delta(B_n(W_T' T)), \tag{4}$$

where $\tilde{T} \in \mathbb{R}^D$ is the textual features projection, W_T is the weight matrix of linear layer, B_n is a batch normalization, and δ is a ReLU activation function. For sake of completeness, we compare various text encoders in Section IV-B2.

C. THE PROPOSED MCA FOR FEATURE-LEVEL FUSION

The core of the framework shown in Fig. 1 is the proposed MCA block. Attention [50] has been a significant breakthrough in several fields, such as natural language processing and computer vision. One type of attention mechanism, known as channel attention, was introduced to increase the network sensitivity to informative features [51]. We exploit a new utilization of channel attention blocks as a fusion technique among several modalities, through which channels are perceived as the modalities distinct representation. The goal of the proposed MCA block is learning the interdependencies for the modalities. It performs modality weighing by emphasizing on the more informative modality to learn a common discriminative representation. In Figure 2, the architecture of the proposed MCA is illustrated. The input to MCA block is the concatenation of both visual $\tilde{V}$ and textual $\tilde{T}$ features projection, which can be mathematically represented as:

$$F = \tilde{V} \mid \tilde{T}, \tag{5}$$

where $\mid$ means channel level concatenation, $F \in \mathbb{R}^{(D\times 1\times C)}$ represents the fused features and C is the number of modalities which is two channels (Visual and Textual). Both modalities share the same size of features dimension (D), where D is set to 1000, to ensure that all modalities are given equal consideration during the fusion procedure.

Inspired by squeeze and excitation (SE) block [51], the proposed MCA block collects features from both modalities and outputs a joint global representation of these features.It then assigns ingenious attention weights to this joint representation to help the model prioritize what crucial features to focus on across all modalities. The aggregation of spatial features into a joint global channel descriptors is employed through employing a global average pooling (GAP) technique as follow:

$$\tilde{F} = \frac{1}{D}\sum_{j=1}^{D} F(j). \tag{6}$$

Subsequently, the process of assigning attention weights to these joint representations is performed through the utilization of two fully connected layers. The first layer incorporates a ReLU activation function, while the second uses a sigmoid activation function, allowing the scoring of each multimodal channel (α) to be evaluated as follows:

$$\alpha = \sigma(W'_{F2}(\delta(W'_{F1}\tilde{F}))), \tag{7}$$

where α, $\tilde{F}$, W_{F1}, and $W_{F2} \in \mathbb{R}^{(1\times 1\times C)}$, σ is a sigmoid activation function and δ is a ReLU activation function. The output of the MCA block is then obtained by scaling to the fused features (F) as follows:

$$F_A = \alpha \otimes F, \tag{8}$$

where $\otimes$ is an element wise multiplication, α represents channel attention weights, and $F_A \in \mathbb{R}^{(D\times 1\times C)}$.

Lastly, the output of the MCA block undergoes processing through a linear layer. This results in the final classification scores of the fused image and text modalities for each class, denoted as y_{MCA}.

Additionally, we investigate alternative attention-based blocks that draw inspiration from channel attention blocks other than SE block [51], specifically Gated Channel Transformation block (GCT) [52], Efficient Channel Attention (ECA) [53], and Selective Kernel (SK) [54]. These blocks are modified to accommodate the integration of multiple modalities, as further discussed in section (IV-B5).

D. SCORE-LEVEL FUSION

Following the classification scores of image (y_V), text (y_T), and feature-level fused MCA (y_{MCA}), the output classification result is determined through an ensemble of the three outputs [55]. The proposed DMCC framework is the ensemble of three pathways described in Eq. 9 as the summation of the three scores of each class.

$$y = y_V + y_T + y_{MCA}. \tag{9}$$

IV. EXPERIMENTAL RESULTS

In order to evaluate the effectiveness of the proposed DMCC framework, extensive experiments are conducted on two publicly available multimodal crisis-related datasets. In this section, first datasets and the implementation settings are presented. Then, we conduct extensive ablation study to gain a better understanding of each component of DMCC framework. Next, quantitative and qualitative analysis of the framework are provided. Finally, DMCC framework is compared against the state-of-the-art methods.

A. DATASETS AND IMPLEMENTATION SETTINGS

Here, the datasets and the implementation settings used in the experiments conduction are discussed.

1) DATASETS

Very few multimodal crisis datasets are publicly available for disaster response classification tasks of social media data. Our research uses the only two available multimodal crisis datasets: CrisisMMD [21] and DMD [17] datasets. Both datasets consist of image-tweet pairs with their annotated class label.

The **CrisisMMD** is a multimodal dataset collected by Alam *et al.* [21] in 2017 during seven different natural disasters. The dataset consists of 18802 samples of tweet image-text pairs, split into 70 %, 15 % and 15 % for train, validation, and test sets respectively. In this study, we adhered to the dataset settings outlined by the dataset authors [20]. Specifically, we selected the samples which the labels of both text and image pairs align for a given task.[1] The dataset comprises three main tasks:

Task 1 Informative and Non-informative: The aim of this task is to evaluate the usefulness of a certain tweet text or image that was collected during a disaster event for humanitarian assistance and aid purposes.

Task 2 Humanitarian Categories: The aim of this task is to understand the type of the crucial and potentially actionable information shared on either tweet image or text, and categorize it into eight classes, which are: (1) Infrastructure and utility damage, (2) Vehicle damage, (3) Rescue, volunteering, or donation efforts, (4) Affected individuals, (5) Injured, or dead people, (6) Missing, or found people, (7) Other relevant information, (8) Not humanitarian.

Task 3 Damage Severity: The purpose of this task is to determine the severity of the damage shown in a tweet image and divide it into three classes: severe, mild, and little to none. This task is an image-based classification task; therefore, we only consider the first two multimodal tasks as in [20].

The **Damage Identification Multimodal Dataset (DMD)** is a benchmark multimodal damage dataset created by Mouzannar et al. [17], which includes damage-related images along with their associated tweets and annotations. The dataset was created to be descriptive of the damage as to direct the most effective resources for each situation. The dataset consists of 5831 captioned images and is split into 4-folds to perform cross validation. In addition, each fold is divided into 75 %, 7 %, and 18 % for training, validation, and testing, respectively. This dataset contains the following six different categories of disaster tweet image-text pairs: (1) Damaged utility and infrastructure: includes damaged buildings, wrecked cars, and destroyed bridges, (2) Damaged nature: includes landslides, avalanches, and falling trees, (3) Fires: includes wildfires and building fires, (4) Floods: includes city, urban and rural, (5) Human damage: includes injuries and deaths, (6) Non damage.

2) IMPLEMENTATION SETTINGS

In our experiments, SGD optimizer was used with a base learning rate of 0.0002 and 0.1 reduction factor when validation loss was saturated. The batch size was 12 and 8 for

[1] The CrisisMMD dataset with the multimodal agreed labels annotations are available at https://crisisnlp.qcri.org/crisismmd

unimodal and multimodal models, respectively. Additionally, the model's training was performed for 80 epochs. Experiments were executed on a machine with an Intel Core i7-9700 CPU with 8 cores and a Nvidia GeForce GTX-1080Ti GPU. Models were implemented using pytorch. The most used performance metrics for classification, namely accuracy, precision, F1 score, and recall were adopted in this study.

B. ABLATION STUDY

We conduct an ablation study to gain insights into the effect of different configurations of DMCC components. All ablation experiments are carried out on the two tasks of CrisisMMD dataset, specifically, task 1 (Informative vs Non-Informative) and task 2 (Humanitarian categories).

1) THE CHOICE OF IMAGE ENCODER NETWORK

The choice of image encoder is crucial, as it governs the effectiveness of the visual representation. Here, we explore several image classification networks on CrisisMMD dataset: Densent121 [42], ResNet101 [41], RegNet [56], Vision Transformers (VIT) [57], and EfficientNetV2 [49]. The results are summarized in Table 1. We infer from the table that EfficientNetV2 achieves the highest recognition accuracy 77.77 % and 74.00 % in task 1 and 2, respectively. These results promote EfficientNetV2 as a good candidate for image encoding.

TABLE 1. Accuracy performance comparison among different image classification networks on CrisisMMD dataset.

Network	Accuracy %	
	Task 1	Task 2
Densenet121	75.23	72.10
Resenet101	76.00	72.15
RegNetY	75.88	73.20
VIT	77.44	71.88
EfficientNetV2	**77.77**	**74.00**

2) THE CHOICE OF TEXT ENCODER NETWORK

Towards finding a descriptive text representation, two text encoders are evaluated, specifically BERT [43] and RoBERTa [58]. The experimental results tabulated in Table 2, indicate BERT has a higher recognition accuracy than RoBERTa by around 7 % and 13 % in task 1 and 2, respectively. These results imply the effectiveness of BERT as a text encoder.

TABLE 2. Comparing the performance of RoBERTa and BERT on CrisisMMD dataset.

Network	Accuracy %	
	Task 1	Task 2
RoBERTa	78.75	60.10
BERT	**85.90**	**73.40**

3) THE IMPACT OF TRANSFER LEARNING FROM TASK 2 TO TASK 1

We investigate the significance of domain specific transfer learning to the attained performance in task 1. Here, we compare two strategies. One strategy is to fine-tune the network starting from the original pretrained weights denoted as "original" in Table 3. The other strategy is fine-tuning the network using weights obtained from task 2 as illustrated in Algorithm 1. This strategy is denoted as "Task2" in Table 3. From the table, it is inferred that using task 2 as a pretrained model yields a higher performance than using the original pretrained model by around 2.5 % and 1 % for text and image modalities, respectively. The performance improvement is due to pretraining on a source domain similar to the target domain.

Algorithm 1 Transfer Learning From Task 2 to Task 1

```
1: Train the network on the dataset for task 2.
2: Utilize the obtained weights as an initialization.
3: Fine-tune the network for task 1.
```

TABLE 3. The effect of transfer learning from task 2 to task 1 on CrisisMMD dataset.

Network	Accuracy %	
	Original	Task2
EfficientNetV2	77.77	**80.00**
BERT	85.90	**86.90**

4) THE CHOICE OF LOSS FUNCTION

Next, we assess the choice of the appropriate loss function for the MCA block, we explore two losses namely: Cross Entropy Loss (CE) and Focal Loss (FL) [59]. Table 4 shows a comparative study between the model performance with CE and FL with different focusing parameter (γ). The results demonstrate that CE has a better performance than FL in task 1. On the other hand, FL with $\gamma = 0.5$ in task 2 performs better than CE by around 1 %. Therefore, CE and FL with $\gamma = 0.5$ are chosen in task 1 and 2 throughout the experiments, respectively.

TABLE 4. Accuracy performance comparison for MCA output among different Loss Functions on CrisisMMD dataset.

Network	Accuracy %	
	Task 1	Task 2
CE	**91.40**	84. 08
FL($\gamma = 0.5$)	90.94	**85.03**
FL($\gamma = 1$)	91.13	83.35
FL($\gamma = 1.5$)	90.8	83.04

5) CHANNEL ATTENTION BLOCKS FOR MULTIMODAL FUSION

To demonstrate the advantage of utilizing squeeze and excitation block as the inspiration to the proposed MCA, we conduct a comparison between MCA and various channel attention blocks specifically: GCT [52], ECA [53], and SK [54] as a multimodal fusion technique between visual and textual modalities. The results are reported in Table 5. As observed, both the MCA and SK blocks demonstrate superior performance compared to the other blocks in task 1. However, the SK block is computationally expensive [54], yet exhibits comparable performance. Moreover, MCA block achieves the highest accuracy in task 2. This results justify the effectiveness of the proposed MCA block.

TABLE 5. Accuracy performance comparison among different channel attention blocks for multimodal fusion on CrisisMMD dataset.

Network	Accuracy %	
	Task 1	Task 2
GCT	91.13	84.08
ECA	90.60	85.00
SK	**91.50**	84.50
MCA (Ours)	**91.40**	**85.03**

C. QUANTITATIVE AND QUALITATIVE ANALYSIS

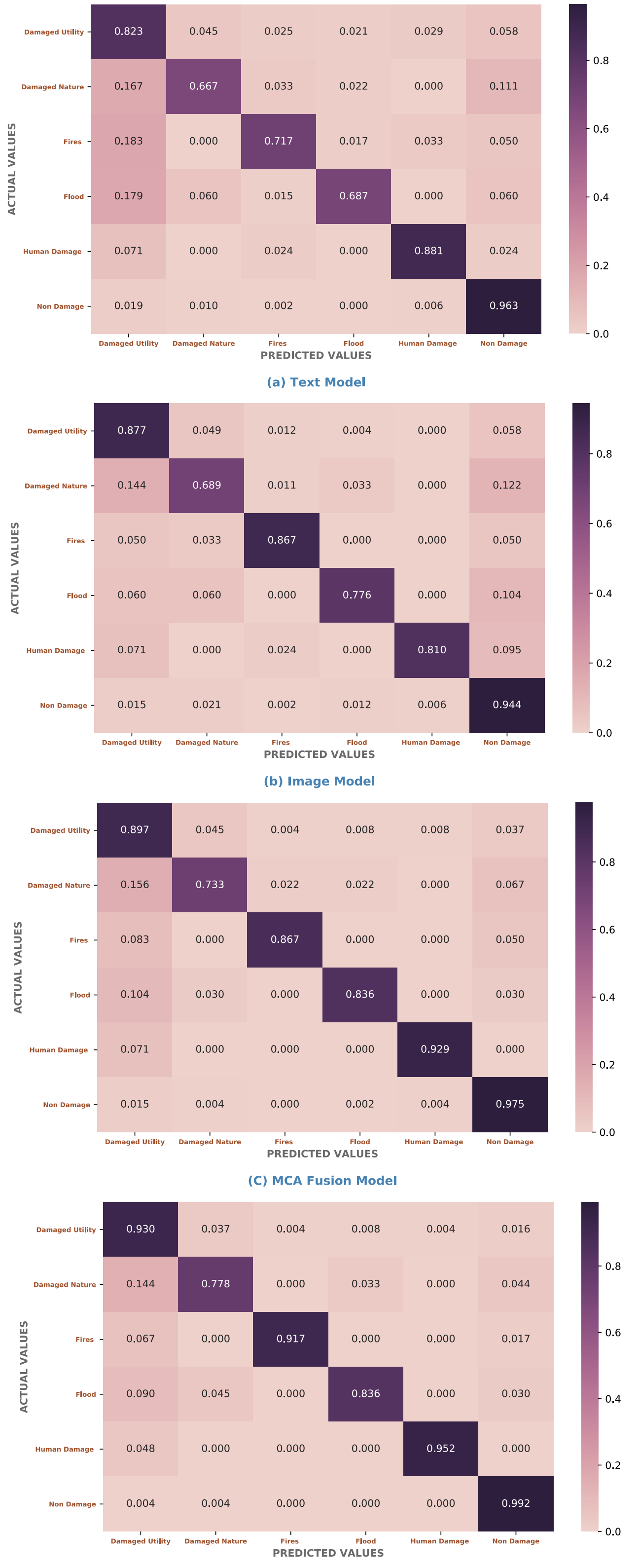
In this section, quantitative and qualitative analysis on DMD datatset are presented, providing insights into the proposed DMCC framework.

1) QUANTITATIVE ANALYSIS

To gain insights about DMCC framework, confusion matrices of unimodal and multimodal models on DMD dataset are depicted in Fig. 3. The figure implies that unimodal models (EfficientV2/BERT) suffer a noticeable confusion among some classes. For example, ''Damaged Nature'' is highly frequently confused with ''Damaged Infrastructure and Utility'' and ''Non-Damage''. This is evidenced by the relatively low true positive rate in ''Damaged Nature'' of 66.7 % and 68.9 % for image and text models, respectively. On the other hand, multimodal models attain a higher true positive rate than single modality; as demonstrated by the 73.3 % true positive rate achieved by MCA model and further enhanced by the proposed DMCC framework, reaching 77.8 %. Conversely, the proposed multimodal models (MCA model and DMCC framework) show a significant decrease in the false positive rate compared to unimodal models. Specifically, when the ''Damaged Nature'' class is mistakenly classified as ''Non-Damage'', the false positive rate drops from 12.2 % and 11.1 % in image and text models, respectively, to 6.7 % in MCA model, and further to 4.4 % in the proposed DMCC framework.

2) QUALITATIVE ANALYSIS

For qualitative analysis, we present in Fig. 4, examples of image and text pairs along with their predictions. Careful observation shows the effectiveness of the proposed DMCC framework over unimodal models (EfficientV2/BERT).



(a) Text Model

(b) Image Model

(C) MCA Fusion Model

(d) Proposed DMCC Model

FIGURE 3. DMD confusion matrices.

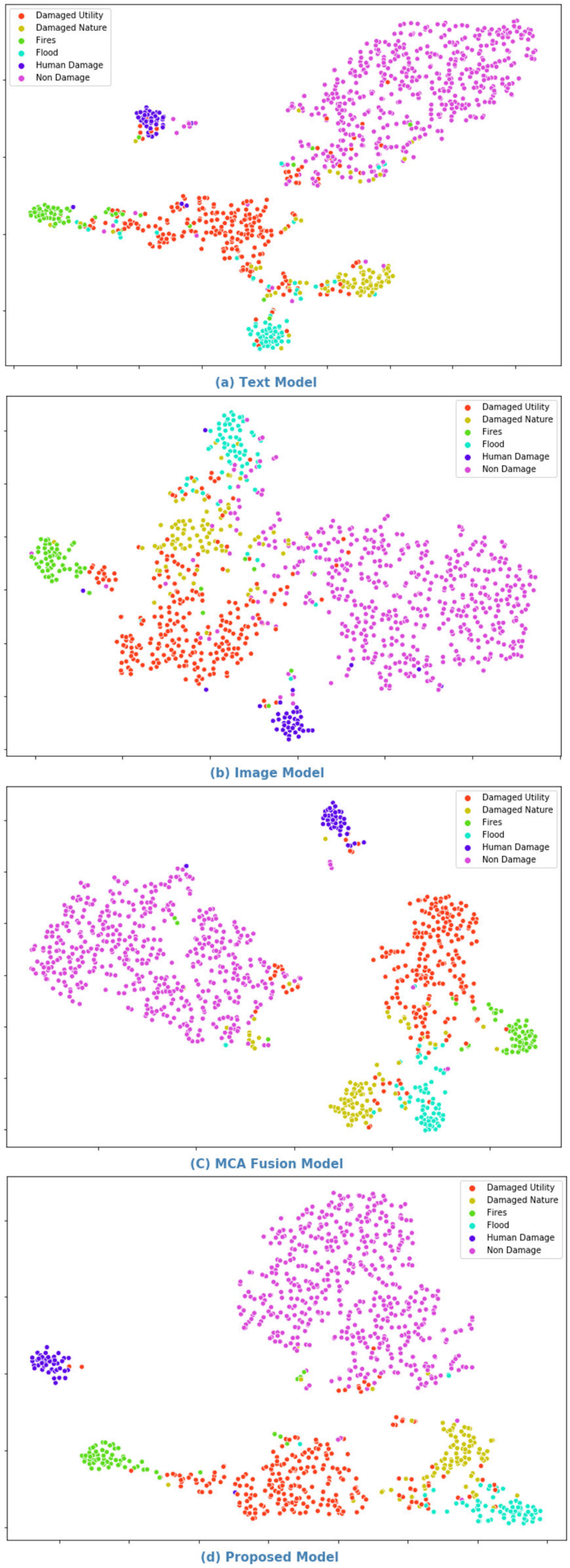
For example, the unimodal models fail in predicting the sample (1). The visual model is not able to discern

Sample	Image-Text Pair	Unimodal Prediction	MCA Fusion Prediction	Proposed DMCC Model Prediction
(1)	 5years ago today a day I will never forget. #HurricaneSandy #BeforeAndAfter #RockawayBeach	Non Damage × Damaged Utility ×	Flood ✓	Flood ✓
(2)	 #fordracing #ford #cortina #platina #rusty #mothernature #red #trees #forrest #growing #rustic #antik #tearaway #volvo #amazon #oldvolvo #oldvolvosneverdietheyjustgetfaster #broken #familycar #graveyard #wreckedcar	Damaged Utility ✓ Damaged Nature ×	Damaged Utility ✓	Damaged Utility ✓
(3)	 Please allow us to cross the border ....#assadcrimes #isiscrimes #poor #poorsyrians #genocide #syrianorphans #syriangenocide #shameonhumanity #shameonthisworld #pain #refugees #syrianrefugees	Non Damage × Human Damage ✓	Human Damage ✓	Human Damage ✓
(4)	 SOMALIA: At least 20 dead and many injured after a deadly truck bomb explodes in the Somalian Capital of Mogadishu, The truck bomb detonated outside a busy hotel in the capital, police expect the death toll to rise, No group has yet claimed responsibility but itâ€™s believed Al-Shabab carried out the attack #Somalia #truckbombing #attack #terrorattack #terrorism	Fires ✓ Damaged Utility ×	Damaged Utility ×	Fires ✓

FIGURE 4. DMD Tweet Image-Text pair Examples with unimodality and multimodal predictions. The symbol (✓) and (×) indicates the correct and incorrect prediction respectively.

the damage caused by the flood and classified the event as "Non Damage". Similarly, the textual model failed, as well, to give a proper prediction and falsely classifies the event as "Damaged Utility". However, the MCA model and the proposed DMCC framework correctly predict the event as "Flood". These results imply the influence of discriminative representation of MCA block on the prediction.

From samples (2) and (3), we notice that either visual or textual model misclassifies the event. Interestingly, MCA



(a) Text Model

(b) Image Model

(C) MCA Fusion Model

(d) Proposed Model

FIGURE 5. DMD features representations.

model exhibits the capability of matching the modality with the correct prediction.

Another noteworthy example is sample (4), where the visual model correctly interprets the picture as "Fire". Conversely, the textual model presumes the text as "Damaged Utility". Additionally, the MCA model incorrectly recognize the sample as "Damaged Utility". This is because the sample includes both fired and damaged vehicles, which may be misclassified as "Fire" and "Damaged Utility". However, the proposed DMCC framework predicts the event correctly as "Fire". This demonstrates the superiority of the DMCC framework in identifying the damage information.

Additionally, we visualize the feature subspace of unimodal (EfficientV2/BERT) and multimodal models (MCA model/DMCC framework) on DMD dataset using t-SNE [60] in Fig. 5. From the figure, we observe that the between class distance is small and the classes are smeared in the unimodal models. On the other hand, similar classes are more distinctly separated in the multimodal models, while the samples in the same class are more compactly clustered. Since the MCA block helps in separating the space in more distinctive manner, we believe that the application of the proposed DMCC framework may help in improving the performance in other application areas as well.

D. COMPARISON AGAINST STATE-OF-THE-ART

Beyond analyzing each component in the proposed DMCC framework, shown in Fig. 1, we also assess the performance of DMCC framework against unimodal baselines (visual/textual) and other state-of-the-art methods. We conduct the experiments on two publicly available multimodal social media crisis datasets: CrisisMMD and DMD. We opt to use EfficientNetV2 (visual) and BERT (textual) as strong baseline to assess DMCC framework.

1) CrisisMMD DATASET

Table 6 presents the results on the CrisisMMD dataset. The MCA model demonstrates a significant improvement over EfficientNetV2, with an average of improvement of ~11 % across all metrics for both tasks. Also, it outperforms BERT with an average of ~4.5 % and ~11 % for tasks 1 and 2, respectively. Compared with the other state-of-the-art multimodal methods namely Ofli et al. [20] and Abavisani et al. [19],[2] the proposed MCA model shows a substantial superiority. Additionally, the proposed DMCC framework surpass the current state-of-the-art methods by ~4 % in task 1 and ~5 % in task 2.

2) DMD DATASET

The DMCC framework evaluation results on DMD dataset are listed in Table 7. The results show that the MCA model has higher recognition accuracy than the unimodal methods by ~5 %. DMCC framework exceeds the

[2] Results in the Table 6 are obtained from running the implementation available at https://github.com/PaulCCCCCCH/Multimodal-Categorization-of-Crisis-Events-in-Social-Media

TABLE 6. CrisisMMD dataset performance comparison of different unimodal and multimodal models on the test set. Here, Acc, P, R, and WF1 denotes the accuracy, precision, recall, and weighted F1-score, respectively.

Approach	Network	Task 1				Task 2			
		Acc.%	R%	P%	WF1%	Acc.%	R%	P%	WF1%
Visual	EfficientnetV2	80.00	80.00	82.80	80.52	74.00	74.03	74.98	72.83
Textual	BERT	86.90	86.90	86.98	86.46	73.40	73.40	79.50	74.10
Multimodal	Ofli *et al.* [20]	84.40	84.00	84.10	84.20	78.40	78.50	78.00	78.30
	SSE-Cross-BERT-DenseNet [19]	88.46	88.46	88.42	88.22	82.67	82.67	83.22	82.64
	MCA (Ours)	91.40	91.40	91.33	91.34	85.03	85.03	85.35	84.94
	DMCC (Ours)	**92.24**	**92.24**	**92.23**	**92.24**	**88.00**	**87.95**	**87.80**	**87.72**

TABLE 7. DMD dataset performance comparison of different unimodal and multimodal models on the test set. STD means standard deviation.

Approach	Network	Accuracy (mean±STD) %
Visual	EfficientnetV2	86.71 ± 1.25
Textual	BERT	85.93 ± 0.82
Multimodal	Mouzannar *et al.* DFMC with SVM [17]	92.62 ± 0.89
	Mouzannar *et al.* FFMC with ANN [17]	92.60 ± 0.77
	MCA (Ours)	91.60 ± 0.60
	DMCC (Ours)	**93.68 ± 0.46**

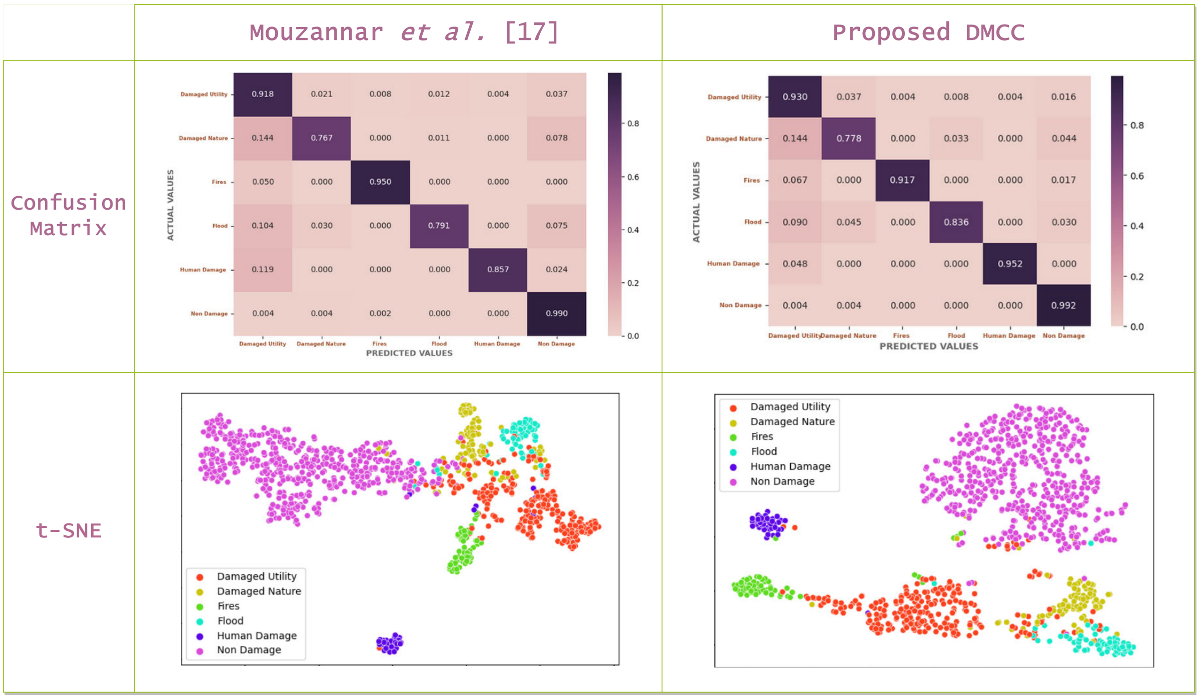
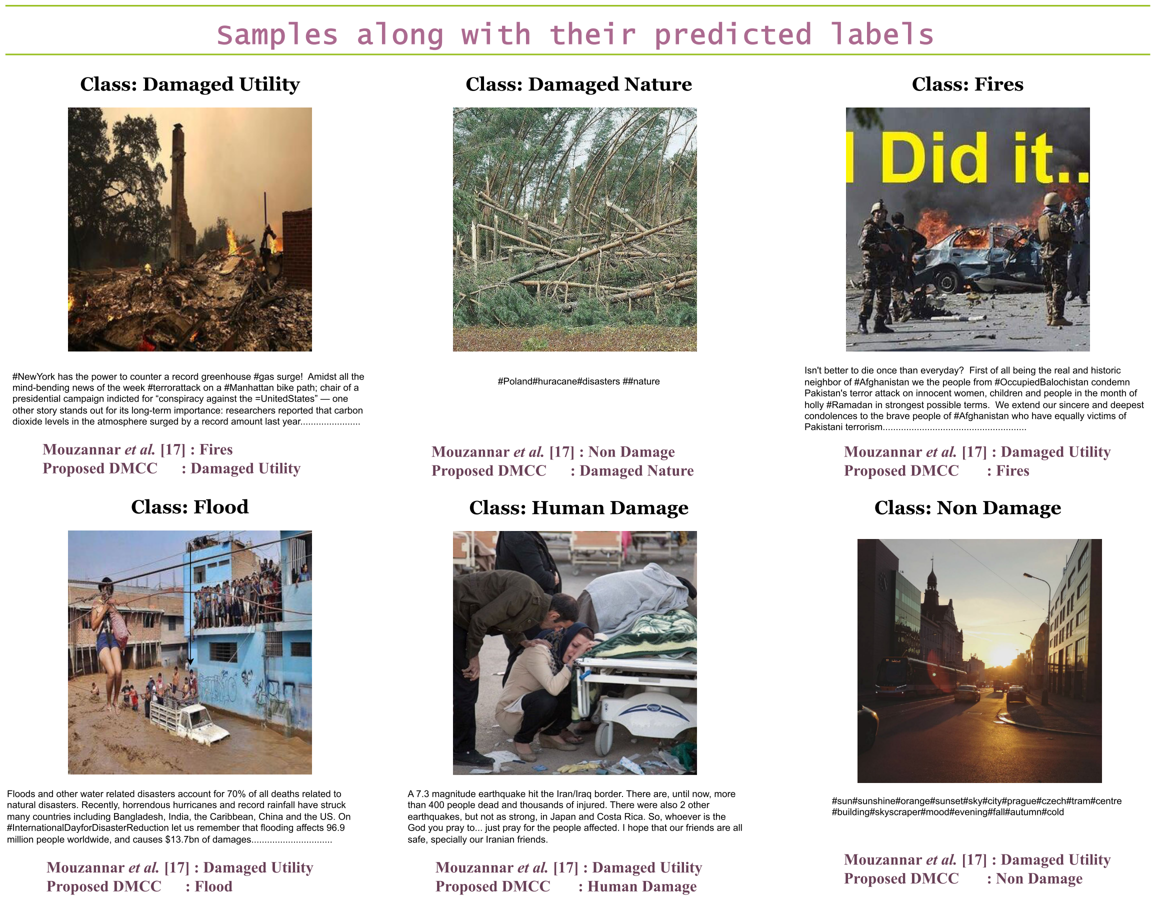


FIGURE 6. SOTA comparison of confusion matrix and t-SNE feature visualization technique for DMD dataset.

state-of-the art method by ∼1 % empowered by the capability of the MCA block in learning a discriminative joint feature representation.

In addition to the comprehensive analysis on DMD dataset presented in Fig. 3, Fig.4 and Fig.5, we extend our evaluation by conducting a detailed comparative analysis between the



#NewYork has the power to counter a record greenhouse #gas surge! Amidst all the mind-bending news of the week #terrorattack on a #Manhattan bike path; chair of a presidential campaign indicted for "conspiracy against the =UnitedStates" — one other story stands out for its long-term importance: researchers reported that carbon dioxide levels in the atmosphere surged by a record amount last year......................

Mouzannar *et al.* [17] : Fires
Proposed DMCC : Damaged Utility

#Poland#huracane#disasters ##nature

Mouzannar *et al.* [17] : Non Damage
Proposed DMCC : Damaged Nature

Isn't better to die once than everyday? First of all being the real and historic neighbor of #Afghanistan we the people from #OccupiedBalochistan condemn Pakistan's terror attack on innocent women, children and people in the month of holly #Ramadan in strongest possible terms. We extend our sincere and deepest condolences to the brave people of #Afghanistan who have equally victims of Pakistani terrorism..

Floods and other water related disasters account for 70% of all deaths related to natural disasters. Recently, horrendous hurricanes and record rainfall have struck many countries including Bangladesh, India, the Caribbean, China and the US. On #InternationalDayforDisasterReduction let us remember that flooding affects 96.9 million people worldwide, and causes $13.7bn of damages..............................

A 7.3 magnitude earthquake hit the Iran/Iraq border. There are, until now, more than 400 people dead and thousands of injured. There were also 2 other earthquakes, but not as strong, in Japan and Costa Rica. So, whoever is the God you pray to... just pray for the people affected. I hope that our friends are all safe, specially our Iranian friends.

Mouzannar *et al.* [17] : Damaged Utility
Proposed DMCC : Human Damage

#sun#sunshine#orange#sunset#sky#city#prague#czech#tram#centre #building#skyscraper#mood#evening#fall#autumn#cold

Mouzannar *et al.* [17] : Damaged Utility
Proposed DMCC : Non Damage

FIGURE 7. **Example tweet text and image pairs showcasing DMD classes, with a focus on comparing the predictions made by the proposed DMCC model and those made by Mouzannar et al. [17] for DMD dataset.**

proposed DMCC framework and the state-of-the-art (SOTA) method proposed by Mouzannar et al. [17][3] in-terms of confusion matrix and t-SNE feature visualization. The confusion matrix presented in Fig. 6 clearly indicates that the DMCC framework consistently outperforms SOTA method across five classes. Particularly, in the human damage class, the DMCC framework exhibits notable improvement compared to the comparative method. In this comparison, the DMCC framework successfully eliminates the confusion with "Non-Damage" and significantly reduces the confusion with "Damaged Infrastructure and Utility" from 11.9 % to 4.8 %. However, it is noteworthy that the method of Mouzannar et al. [17] displays superior accuracy specifically in classifying fire instances, as indicated by the confusion matrix.

Furthermore, Fig. 6 compares t-SNE feature visualizations of DMCC framework and SOTA. It shows large inter-class distance in certain classes, particularly "Non-Damage" and "Human Damage" within the feature subspace, demonstrating that our DMCC framework learns more discriminative feature representation than SOTA.

Additionally, Tweet text and image pairs exemplifying different DMD classes are presented in Fig. 7. DMCC framework classifies these samples correctly, while the SOTA method by Mouzannar et al. [17] demonstrates instances of misclassifications.

V. CONCLUSION

In this paper, a novel DMCC framework for the fusion of multiple modalities in the context of crisis-related posts categorization was presented. Our approach utilized two parallel deep learning networks, specifically EfficientNetV2 and BERT, to extract visual and textual features, respectively. Additionally, a feature-level fusion approach named MCA block was proposed, which effectively fuses multiple modalities by assigning distinct weights to each modality. The MCA block effectively filters out any irrelevant or misleading information and selectively fuses the informative components from each modality. The proposed DMCC framework was quantitatively and qualitatively evaluated on two publicly available datasets: CrisisMMD and DMD. The experimental findings demonstrated that the proposed DMCC framework

[3] Fig. 6 and Fig.7 are produced from running the implementation available at https://github.com/husseinmozannar/multimodal-deep-learning-for-disaster-response

outperformed the current state-of-the-art model by $\sim$4 % (task 1) and by $\sim$5 % (task 2) on CrisisMMD dataset. Additionally, it achieved 93.68 % surpassing the state-of-the-art method by $\sim$1 %. Furthermore, the results of the qualitative analysis were consistent with the quantitative findings, indicating that the inclusion of an attention mechanism yields a notable improvement in the analysis of crisis-related social media data. we believe that the application of the proposed DMCC framework may help in improving the performance in other application areas as well.

REFERENCES

[1] T. Simon, A. Goldberg, and B. Adini, ''Socializing in emergencies—A review of the use of social media in emergency situations,'' *Int. J. Inf. Manage.*, vol. 35, no. 5, pp. 609–619, Oct. 2015.

[2] R. Samuels, J. E. Taylor, and N. Mohammadi, ''Silence of the tweets: Incorporating social media activity drop-offs into crisis detection,'' *Natural Hazards*, vol. 103, no. 1, pp. 1455–1477, Aug. 2020.

[3] X. Zhou and L. Chen, ''Event detection over Twitter social media streams,'' *VLDB J.*, vol. 23, no. 3, pp. 381–400, Jun. 2014.

[4] L. Huang, P. Shi, H. Zhu, and T. Chen, ''Early detection of emergency events from social media: A new text clustering approach,'' *Natural Hazards*, vol. 111, no. 1, pp. 851–875, Mar. 2022.

[5] N. Said, K. Ahmad, M. Riegler, K. Pogorelov, L. Hassan, N. Ahmad, and N. Conci, ''Natural disasters detection in social media and satellite imagery: A survey,'' *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 31267–31302, Nov. 2019.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, ''Tweet analysis for real-time event detection and earthquake reporting system development,'' *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 4, pp. 919–931, Apr. 2013.

[7] P. Burnap, M. L. Williams, L. Sloan, O. Rana, W. Housley, A. Edwards, V. Knight, R. Procter, and A. Voss, ''Tweeting the terror: Modelling the social media reaction to the Woolwich terrorist attack,'' *Social Netw. Anal. Mining*, vol. 4, no. 1, p. 206, Dec. 2014.

[8] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, ''Discovering shifts to suicidal ideation from mental health content in social media,'' in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2016, pp. 2098–2110.

[9] H. A. Schwartz et al., ''More evidence that Twitter language predicts heart disease: A response and replication,'' *PsyArXiv*, Mar. 2018, doi: 10.31234/osf.io/p75ku.

[10] K. Muniz-Rodriguez, S. K. Ofori, L. C. Bayliss, J. S. Schwind, K. Diallo, M. Liu, J. Yin, G. Chowell, and I. C.-H. Fung, ''Social media use in emergency response to natural disasters: A systematic review with a public health perspective,'' *Disaster Med. Public Health Preparedness*, vol. 14, no. 1, pp. 139–149, Feb. 2020.

[11] G. Beigi, X. Hu, R. Maciejewski, and H. Liu, ''An overview of sentiment analysis in social media and its applications in disaster relief,'' *Sentiment Anal. Ontology Eng.*, vol. 639, pp. 313–340, Mar. 2016.

[12] H. Shekhar and S. Setty, ''Disaster analysis through tweets,'' in *Proc. Int. Conf. Adv. Comput., Commun. Inform. (ICACCI)*, Aug. 2015, pp. 1719–1723.

[13] S. E. Middleton, L. Middleton, and S. Modafferi, ''Real-time crisis mapping of natural disasters using social media,'' *IEEE Intell. Syst.*, vol. 29, no. 2, pp. 9–17, Mar. 2014.

[14] S. Cresci, A. Cimino, F. Dell'Orletta, and M. Tesconi, ''Crisis mapping during natural disasters via text analysis of social media messages,'' in *Proc. WISE*, Dec. 2015, pp. 250–258.

[15] K. Kitazawa and S. A. Hale, ''Social media and early warning systems for natural disasters: A case study of typhoon Etau in Japan,'' *Int. J. Disaster Risk Reduction*, vol. 52, Jan. 2021, Art. no. 101926.

[16] M. Bica, L. Palen, and C. Bopp, ''Visual representations of disaster,'' in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2017, pp. 1–15.

[17] H. Mouzannar, Y. Rizk, and M. Awad, ''Damage identification in social media posts using multimodal deep learning,'' in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2018, pp. 1–15.

[18] E. Hossain, M. M. Hoque, E. Hoque, and Md. S. Islam, ''A deep attentive multimodal learning approach for disaster identification from social media posts,'' *IEEE Access*, vol. 10, pp. 46538–46551, 2022.

[19] M. Abavisani, L. Wu, S. Hu, J. Tetreault, and A. Jaimes, ''Multimodal categorization of crisis events in social media,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14667–14677.

[20] F. Ofli, F. Alam, and M. Imran, ''Analysis of social media data using multimodal deep learning for disaster response,'' in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2020, pp. 1–10.

[21] F. Alam, F. Ofli, and M. Imran, ''Crisismmd: Multimodal Twitter datasets from natural disasters,'' in *Proc. 12th Int. AAAI Conf. Web Social Media (ICWSM)*, Jun. 2018, pp. 1–9.

[22] T. Sakaki, M. Okazaki, and Y. Matsuo, ''Earthquake shakes Twitter users: Real-time event detection by social sensors,'' in *Proc. 19th Int. Conf. World Wide Web*, Apr. 2010, pp. 851–860.

[23] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, ''Extracting information nuggets from disaster-related messages in social media,'' in *Proc. Int. Conf. Inf. Syst. Crisis Response Manage.*, 2013, pp. 791–801.

[24] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, ''Practical extraction of disaster-relevant information from social media,'' in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 1021–1024.

[25] J. P. Singh, Y. K. Dwivedi, N. P. Rana, A. Kumar, and K. K. Kapoor, ''Event classification and location prediction from tweets during disasters,'' *Ann. Oper. Res.*, vol. 283, nos. 1–2, pp. 737–757, Dec. 2019.

[26] C. Caragea, A. Silvescu, and H. A. Tapia, ''Identifying informative messages in disaster events using convolutional neural networks,'' in *Proc. Int. Conf. Inf. Syst. Crisis Response Manag.*, May 2016, pp. 137–147.

[27] Y. Kim, ''Convolutional neural networks for sentence classification,'' in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–6.

[28] T. D. Nguyen, K. Al-Mannai, R. S. Joty, H. Sajjad, M. Imran, and P. Mitra, ''Robust classification of crisis-related data on social networks using convolutional neural networks,'' in *Proc. Int. Conf. Web Social Media*, 2017, pp. 632–635.

[29] S. Madichetty and M. Sridevi, ''Detecting informative tweets during disaster using deep neural networks,'' in *Proc. 11th Int. Conf. Commun. Syst. Netw. (COMSNETS)*, Jan. 2019, pp. 709–713.

[30] G. Burel, H. Saif, M. Fernández, and H. Alani, ''On semantics and deep learning for event detection in crisis situations,'' in *Proc. Workshop Semantic Deep Learn. (SemDeep)*, 2017, pp. 1–13.

[31] A. Alharbi and M. Lee, ''Crisis detection from Arabic tweets,'' in *Proc. 3rd Workshop Arabic Corpus Linguistics*, 2019, pp. 72–79.

[32] D. T. Nguyen, F. Ofli, M. Imran, and P. Mitra, ''Damage assessment from social media imagery data during disasters,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Jul. 2017, pp. 569–576.

[33] F. Alam, M. Imran, and F. Ofli, ''Image4Act: Online social media image processing for disaster response,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Jul. 2017, pp. 601–604.

[34] X. Li, D. Caragea, H. Zhang, and M. Imran, ''Localizing and quantifying damage in social media images,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 194–201.

[35] A. Kumar and J. P. Singh, ''Disaster severity prediction from Twitter images,'' *Intell. Enabled Res.*, vol. 1279, pp. 65–73, Dec. 2020.

[36] F. Alam, F. Ofli, M. Imran, T. Alam, and U. Qazi, ''Deep learning benchmarks and datasets for social media image classification for disaster response,'' in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Dec. 2020, pp. 151–158.

[37] F. Alam, T. Alam, F. Ofli, and M. Imran, ''Robust training of social media image classification models,'' *IEEE Trans. Computat. Social Syst.*, early access, Dec. 26, 2022, doi: 10.1109/TCSS.2022.3230839.

[38] Y. Rizk, H. S. Jomaa, M. Awad, and C. Castillo, ''A computationally efficient multi-modal classification approach of disaster-related Twitter images,'' in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, Apr. 2019, pp. 2050–2059.

[39] R. M. Haralick, K. Shanmugam, and I. Dinstein, ''Textural features for image classification,'' *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973.

[40] G. J. Daugman, ''Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters,'' *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 2, no. 7, pp. 1160–1169, 1985.

[41] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[44] 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.

[45] A. Kumar, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, "A deep multimodal neural network for informative Twitter content classification during emergencies," *Ann. Oper. Res.*, vol. 319, no. 1, pp. 791–822, Dec. 2022.

[46] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Global, 2010, pp. 242–264.

[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[48] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (PMLR)*, 2019, pp. 6105–6114.

[49] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn. (PMLR)*, 2021, pp. 10096–10106.

[50] A. Vaswani, M. N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[51] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[52] Z. Yang, L. Zhu, Y. Wu, and Y. Yang, "Gated channel transformation for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11791–11800.

[53] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[54] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[55] Z.-H. Zhou, "Ensemble learning," in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Boston, MA, USA: Springer, 2009, pp. 270–273, doi: 10.1007/978-0-387-73003-5_293.

[56] J. Xu, Y. Pan, X. Pan, S. Hoi, Z. Yi, and Z. Xu, "RegNet: Self-regulated network for image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Mar. 25, 2022, doi: 10.1109/TNNLS.2022.3158966.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

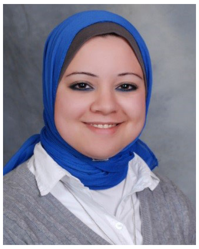
[58] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[59] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[60] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

MARIHAM REZK received the B.Sc. degree (Hons.) in electronics and communications engineering from the Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt, in 2019, where she is currently pursuing the M.Sc. degree in electronics and communications engineering. Since 2019, she has been a Graduate Teaching Assistant with the Department of Electronics and Communication Engineering, AASTMT, and a Research Assistant with the Intelligent Systems Laboratory, AASTMT. Her research interests include machine learning, deep learning, computer vision, and multimodal fusion. She is a Reviewer of IEEE Access.

NOURELDIN ELMADANY received the Ph.D. degree in electrical and computer engineering from Toronto Metropolitan University. From 2020 to 2022, he was with the Vector Institute, Toronto, ON, Canada. He is currently an Assistant Professor with the Department of Electronics and Communications Engineering, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt. His current research interests include machine learning, computer vision and statistical pattern recognition, and deep learning, especially action/activity/event localization and recognition. He was a recipient of the First Place at EPIC Kitchen Challenge, in 2019, and the Visiting Fellowship from Microsoft Research Asia, in 2017.



RADWA K. HAMAD received the B.Sc. (Hons.) and M.Sc. degrees in electronics and communications engineering from the Arab Academy for Science, Technology and Maritime Transport (AASTMT), Alexandria, Egypt, in August 2004 and March 2007, respectively, and the Ph.D. degree in electronics and communications engineering from the Faculty of Engineering, Alexandria University, Alexandria, in 2015. From 2004 to 2015, she was a Teaching and Research Assistant with the Department of Electronics and Communications Engineering, AASTMT, where she is currently an Associate Professor. Her research and teaching interests include channel estimation, direction of arrival, wireless communications, signal and image processing, and 5G antenna design.

EHAB F. BADRAN (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in electrical engineering from Assiut University, Asyut, Egypt, in May 1995 and March 1998, respectively, and the M.Sc. and Ph.D. degrees in electrical engineering from Louisiana State University (LSU), Baton Rouge, LA, USA, in May 2001 and May 2002, respectively. From 1995 to 1998, he was an Instructor with the Department of Electrical Engineering, Assiut University, where he was promoted to an Assistant Lecturer, in May 1998. From January 2000 to May 2002, he was a Teaching and Research Assistant with the Department of Electrical and Computer Engineering, LSU, during the Ph.D. studies. From September 2002 to August 2003, he was an Assistant Professor with the Department of Electrical Engineering, Assiut University. From September 2003 to May 2007, he has an Assistant Professor with the Department of Electronics and Communications Engineering, Arab Academy for Science, Technology and Maritime Transport, Alexandria, Egypt, where he was an Associate Professor, from June 2007 to May 2011. In June 2011, he was promoted to a Professor. His research and teaching interests include wireless communications, signal processing, MIMO systems, image processing, and communication networks.