Crisis-DIAS: Towards Multimodal Damage Analysis - Deployment, Challenges and Assessment

Mansi Agarwal,[1*] **Maitree Leekha,**[1*] **Ramit Sawhney,**[2] **Rajiv Ratn Shah**[3]
[1]Delhi Technological University, New Delhi, India
[2]Netaji Subhas Institute of Technology, New Delhi, India
[3]Indraprastha Institute of Information Technology, New Delhi, India
{mansiagarwal_bt2k16, maitreeleekha_bt2k16}@dtu.ac.in, ramits.co@nsit.net.in, rajivratn@iiitd.ac.in

Abstract

In times of a disaster, the information available on social media can be useful for several humanitarian tasks as disseminating messages on social media is quick and easily accessible. Disaster damage assessment is inherently multi-modal, yet most existing work on damage identification has focused solely on building generic classification models that rely exclusively on text or image analysis of online social media sessions (e.g., posts). Despite their empirical success, these efforts ignore the multi-modal information manifested in social media data. Conventionally, when information from various modalities is presented together, it often exhibits complementary insights about the application domain and facilitates better learning performance. In this work, we present Crisis-DIAS, a multi-modal sequential damage identification, and severity detection system. We aim to support disaster management and aid in planning by analyzing and exploiting the impact of linguistic cues on a unimodal visual system. Through extensive qualitative, quantitative and theoretical analysis on a real-world multi-modal social media dataset, we show that the Crisis-DIAS framework is superior to the state-of-the-art damage assessment models in terms of bias, responsiveness, computational efficiency, and assessment performance.

1 Introduction

Context & Scope. Social media's rapidly increasing ubiquity (Global Digital Report 2019) has made it one of the primary sources for a large multitude of users to engage in mass discussions about disasters and the damage caused by crises (Ahmad et al. 2019). A large number of domains, including public health, economy, and politics (Chew and Eysenbach 2010) (Llorente et al. 2015) (Conover et al. 2013), utilize this gold-mine of data to foster social media research. This helps in analyzing the user's perspective and understanding how user-generated content enhances the user-authority relationship (Reuter and Kaufhold 2018). In times of crisis, users upload a large amount of data in the form of *text, images, video, audio*, etc. which indicate the severity of the impact of the disaster. This data is voluminous, and therefore, there is a need of systems which can automatically identify and highlight a social media post if it indicates severe damage, by analyzing the information present in different media forms.

Challenges. One of the major challenges associated with isolating the media content useful for crisis management (Imran et al. 2015) is the sheer quantity of the social media posts which makes it difficult to identify useful and actionable content in real-time. Therefore, there is a need for automatic systems to identify the presence of damage and analyze it, which could benefit the emergency management process.

While unimodal damage analysis frameworks are efficient, they are not able to assess damage as effectively for social media posts. Text and images are very likely interrelated (Vempala and Preoţiuc-Pietro 2019). Therefore, to develop an effective system we must employ an interdisciplinary approach which can leverage data from different media forms.

Another challenge to the design of emergency assessment systems is how their complexity correlates with their ease of deployment and scalability. A system should be able to perform well in real-time and possess the responsiveness needed to provide aid.

Contributions. Motivated by the cause of *humanitarian* aid in times of disasters, we propose Crisis-DIAS, a novel end-to-end gated multimodal framework that leverages textual and visual cues from user-uploaded information on social media. The system performs a twofold task of first identifying whether a social media post indicates the presence of infrastructural damage, and further assesses the severity level. It does so by extracting cues from text and image modalities, and then *intelligently* merging them by using attention mechanism along with a gated framework to stitch the individual tasks together to serve as an end-to-end service.

Crisis-DIAS is responsive and simple, yet efficient and performs well in real-time multi task damage assessment. Furthermore, we report an extensive comparative quantitative, quantitative and error analysis on a real-world dataset. Our results show how Crisis-DIAS, by learning to combine the best features of different modalities, improves over uni-

[*]Both the authors contributed equally, and wish that they be regarded as joint First Authors.

modal frameworks.

2 Related Work

To develop an effective social media damage assessment tool, we must employ an interdisciplinary approach to assess possible solutions.

Unimodal Disaster Analysis from Text. (Sreenivasulu and Sridevi 2019) analyzed the messages from micro blogs for detecting the *informative* ones, which could be further utilized for other related disaster assessment tasks. They employed a Convolutional Neural Network (CNN) for modeling the text classification problem, using the dataset curated by (Alam, Ofli, and Imran 2018a). On a broader scale, (Alam, Ofli, and Imran 2019) developed an automatic data processing service which analyzed the disaster-related social media text to identify (i) the type of disaster depicted in the message, (ii) whether it was informative and (iii) the type of humanitarian information it contained. They used classical machine learning and deep learning techniques to model text on a dataset formed from several benchmark crisis datasets (Alam, Ofli, and Imran 2018a) (Imran et al. 2014) (Olteanu et al. 2014).

Unimodal Disaster Analysis from Image. (Chaudhuri and Bose 2019) analyzed the images from earthquake-hit smart urban environments to detect human and life damage, using a CNN architecture. (Alam, Ofli, and Imran 2018b) proposed a vision-based damage assessment pipeline designed to (i) filter *relevant* social media images for further assessment and (ii) analyzing the damage severity. However, analysis of disaster content based on unimodal cues may sometimes lead to incorrect predictions.

Multimodal Disaster Analysis. The existing work on combining features from different media sources focuses primarily on categorizing the type of humanitarian damage, without the identification or further assessment of such posts indicating damage. (Rizk et al. 2019) used traditional handcrafted features for text and images from a home-grown dataset and combined them using Support Vector Machine to determine the type of damage indicated in a Twitter post. In another work, (Pouyanfar et al. 2019) used Multiple Correspondence Analysis to fuse audio and visual features to classify what *concept* (*humanitarian category*) the posts indicate. (Mouzannar, Rizk, and Awad 2018) used textual and visual cues to categorize the type of damage indicated in a post, using decision and feature fusion.

We compare this existing use of decision and feature with another advanced fusion technique using attention mechanism. Attention fusion not only performs better than the former two approaches but also learns which features are the most *important* ones and accordingly attends them. Further, we use state-of-the-art computer vision techniques such as deep CNNs to extract spatial information from the images, alongside recurrent networks and advanced embeddings to model the text.

3 Problem Definition and Dataset

Recently, several datasets on crisis damage analysis have been released to foster research in the area (Said et al. 2019). In this work, we have used the first multimodal, labeled, publicly available damage related Twitter dataset, CrisisMMD, created by (Alam, Ofli, and Imran 2018a). The dataset was collected by crawling the blogs posted by users during seven natural disasters, including floods, wildfires, hurricanes and earthquakes. It is hierarchical, *i.e.*, the class labels at each stage depend on the annotation of the previous stage.

The proposed system in Figure 1 aims to solve the three humanitarian tasks formulated below using this dataset.

- **Task 1- Informativeness.**
  (Alam, Ofli, and Imran 2018a) defines a tweet as *informative* if it serves to be useful in identifying areas where damage has occurred due to disaster and provides separate labels for text and images.

  Let F_{info} be the binary valued function, which takes in a multimodal tweet and maps to either 1 or 0, depending on whether the tweet is *informative* or not. More precisely, let M_x be a multimodal tweet in the CrisisMMD dataset (D), having text t_x and image i_x. Then, $\forall M_x \in D$, $F_{info}(t_x, i_x) = 1$ if M_x is *informative*, and 0 otherwise.

  As both text and images have separate annotations, we begin by estimating separate functions for the two modalities and then combine them. We modify the existing three class labels (belonging to the domain $\{$*informative, non-informative, none*$\}$) by combining *non-informative* and *none* classes into one class- calling it *non-informative*, as neither helps in further damage assessment.
  Class Distribution:
  Text- Informative (12877) : *Non-Informative (5249)*
  Image- Informative (9375) : *Non-Informative (8751)*

- **Task 2- Infrastructural Damage.**
  Crisis-DIAS aims to identify and assess the damage in the tweets for severity. The damage in the *informative* tweets from Task 1 may be of many different kinds (Alam, Ofli, and Imran 2018a; Alam et al. 2018). However, for the rescue operation groups to provide aid, it makes sense to focus on only those tweets which signify physical damage, or where people might be stuck (Alam, Ofli, and Imran 2018a).

  The CrisisMMD dataset identifies several humanitarian categories for damage, namely- *Infrastructure and utility damage, Vehicle damage, Affected individuals, Missing or found people, Other relevant information*, *None*. Out of these categories, the tweets belonging to *infrastructure and utility damage* suffer from physical damage, such as broken structures, etc. Although other categories, such as *missing or found people* are also informative for assessment, but not for analyzing the severity.

  Therefore, Task 2 for Crisis-DIAS involves analyzing the tweets for *infrastructural and utility damage* in them. This task forms a bridge between the other two, which separately aim to identify and assess the damage, respectively.

  Given a multimodal tweet $M_x \in D$, with text t_x and image i_x, the objective for Task 2 is to identify whether it suffers from *infrastructural* damage or not.

Specifically, function F_{infra} must be such that, $F_{infra}(t_x, i_x) = 1$, if the tweet M_x is *informative* and has *infrastructural damage*, and 0 otherwise.

That is, Task 2 must ideally be only for the *informative* tweets from Task 1. However, as we propose an end to end system *i.e.*, given a tweet, it will analyze the tweet for both damage identification, and severity. Hence, we must be able to discard the *non-informative* posts which do not need any further analysis. Moreover, there may always be a chance that a *non-informative* tweet is misclassified as *informative*. Since such data points do not have any label for Task 2, therefore, to allow the system to handle all the cases we slightly modify and augment the existing label set in CrisisMMD. As per the binary problem formulation, we take the class label as *non-infrastructural damage* for all samples not in the *infrastructural and utility damage* class. This also includes the *non-informative* samples from Task 1, which originally had no label for Task 2.

As in Task 1, we have separate labels for text and image modalities, which we take leverage of by first, estimating separate functions for each and then, combining them.
Class Distribution:
Text- Infrastructural damage (1428) : *Non-Infrastructural damage (16698)*
Image- Infrastructural damage (3624) : *Non-Infrastructural damage (14502)*

- **Task 3- Severity Assessment.**
Given a multimodal tweet $M_x(t_x, i_x)$ suffering from infrastructural damage (*i.e.*, $F_{infra}(t_x, i_x) = 1$), Task 3 aims to analyze the severity level of damage in three broad categories namely- *no, mild* and *high* damage. However, as reasoned for Task 2, we perform the task on all the tweets, even the samples classified as *non-infrastructural damage* in Task 2 by labelling them as having *no damage*, to support the end-to-end framework. (Alam, Ofli, and Imran 2018a) provides only one tweet level label for severity analysis, which we learn by leveraging from the multimodal cues.
Class Distribution:
No (15068) : *Mild (842)* : *Severe (2216)*

4 Crisis-DIAS

Figure 1 illustrates the system-level diagram for Crisis-DIAS. This section describes in depth the various techniques used in modeling each task, and the way these tasks are stitched together to form an end-to-end multimodal service.

4.1 Pre-Processing

Image Pre-Processing. The images were resized to 299x299 for the transfer learning (Yosinski et al. 2014) model and then normalized in the range [0,1].
Text Pre-Processing. All *http* URLs, retweet headers of the form *RT*, punctuation marks, and twitter user handles specified as *@username* were removed and further lemmatized.

4.2 Vision Pipeline

For the proposed Crisis-DIAS framework, we used the Inception-v3 model (Szegedy et al. 2016), pre-trained on the ImageNet Dataset (Deng et al. 2009). The same architecture has been used for modeling the three tasks, where only the softmax layer changes as per the number of labels.

4.3 Linguistic Pipeline

For the proposed framework, we use Recurrent Convolutional Neural Network (RCNN) (Lai et al. 2015) as the text classification model. It adds a recurrent structure to the convolutional block, therefore capturing contextual information with long term dependencies, and the phrases which play a key role at the same time.

4.4 Multimodal Fusion

The availability of data from different media sources has encouraged researchers to explore and leverage the potential boost in performance by combining unimodal classifiers trained on individual modalities (Asvadi et al. 2017) (Wang, Gong, and Li 2017). We experiment with three different fusion techniques to unify text and image unimodal classifiers for Task 3, namely Decision Fusion, Feature Fusion, and Attention Fusion. The classifiers for Tasks 1 and 2 have been merged with a slight difference.

- **Decision Fusion.** In Decision Fusion, the softmax outputs of the linguistic and vision pipelines are combined in a weighted fashion. The final prediction is given by the class with the maximum weighted score. As a baseline, we use equal weights for both the modalities. To improve this baseline, we use the F1 scores of the two unimodal models to scale their outputs.
- **Feature Fusion.** In case of Feature Fusion, we take the outputs from the last layer before the softmax layer of the two pipelines as inputs to a meta-classifier. We propagate the concatenated tensor of features to a dense and a softmax layer, which gives us the class probability distribution. The optimal weights for these layers were found by training with the weights of all the other layers held constant.
- **Attention Fusion.** The idea of attention fusion is to attend particular input features as compared to others while predicting the output class. Figure 2 illustrates the implementation of this attention mechanism (Hua and Zhang 2004) (Vaswani et al. 2017). The features from the linguistic and vision pipelines are concatenated in the same way as in feature fusion. This is followed by a softmax layer to learn the attention weights for each feature dimension, i.e., the attention weight α_i for a feature x_i is given by:

$$\alpha_i = softmax(\sum_{j=1}^{p} W_{ji} * x_j) \tag{1}$$

$$= \frac{exp(\sum_{j=1}^{p} W_{ji} * x_j)}{\sum_{i=1}^{p} exp(\sum_{j=1}^{p} W_{ji} * x_j)} \tag{2}$$

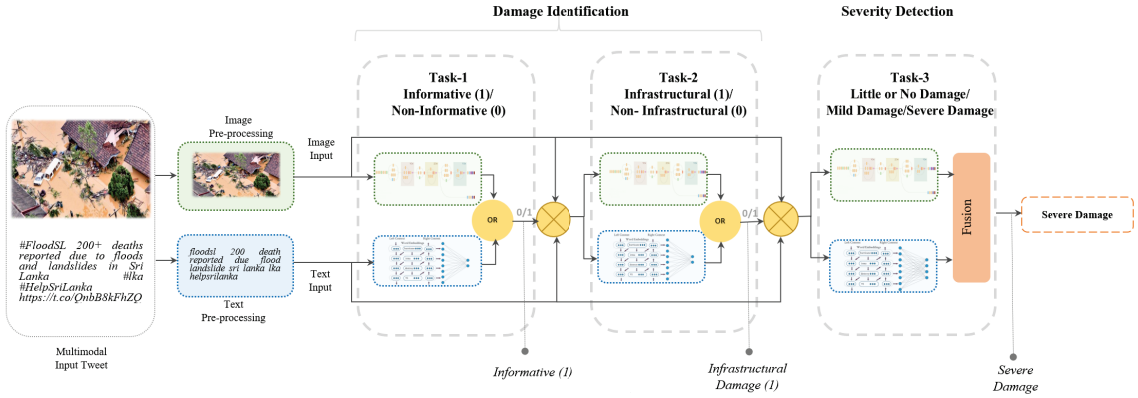

Figure 1: Crisis-DIAS Architecture.

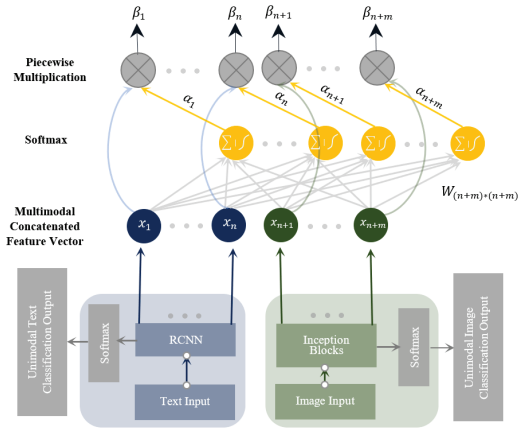

Figure 2: Attention mechanism for multimodal fusion.

Therefore, the input feature after applying the attention weights is,

$$\beta_i = \alpha_i * x_i \tag{3}$$

where, $i, j \in 1, 2, .., p$ and p is the total number of dimensions in the multimodal concatenated feature vector. W is the weight matrix learned by the model.

This vector of attended features is then used to classify the given multimodal input. With this type of fusion, we can also analyze how the different modalities are interacting with each other employing their attention weights.

4.5 Gated Multimodal Architecture

We propose a gated multimodal framework to combine the models for the three tasks together, as shown in Figure 1. Given the hierarchical nature of the dataset, the output for one task regulates the output for the next task, in a sequential fashion. We achieve this by treating the outputs of the tasks as *binary gates*.

For Task 1, the output from the linguistic and vision pipelines is 1 for *Informative* samples, and 0 otherwise. We combine the two outputs by performing a logical OR operation ($\oplus$), i.e., the sample is *Informative* if either the text or the image modality is predicted as being *Informative* by the respective models, *i.e.,* the combined label (τ_1) is,

$$\tau_1 = L_1 \oplus I_1 \tag{4}$$

where L_1 and I_1 are the outputs of the linguistic and vision pipelines of Task 1, respectively.

Similarly, for Task 2, we combine the outputs of the two pipelines (L_2 and I_2) to give τ_2. However, the final output for Task 2 is also dependent on the output of Task 1 (τ_1) and therefore, we use τ_1 as a multiplicative gate for the inputs to the linguistic and vision pipelines (image pixel arrays and token-indexed tweets) of Task 2. More precisely, if τ_1 is 1 (*Informative*), we multiply the model inputs by 1, and the sample is *considered* for Task 2. If τ_1 is 0 (*Non-Informative*), the inputs are multiplied by 0 and the prediction by Task 2 is 0 (*Non-Infrastructural Damage*) always. Therefore,

$$\tau_2 = \begin{cases} L_2 \oplus I_2, & \text{if } \tau_1 = 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

For Task 3, the final output from the fusion layer must also be regulated by the output from Task 2; and therefore, τ_2 is used as a gate. If τ_2 is 1 (*Infrastructural Damage*), the inputs to the vision and linguistic pipelines are multiplied by 1, and the sample is *considered* for severity assessment in Task 3. But if τ_2 is 0 (*Non-Infrastructural Damage*), the inputs for the pipelines are multiplied by 0 and the output for Task 3 is always 0 *i.e. Little or No Damage*. More precisely,

$$\tau_3 = \begin{cases} AttentionFusion(L_3, I_3), & \text{if } \tau_2 = 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Since we have modified the annotations to have labels for all samples on all the tasks, the proposed architecture can catch the samples incorrectly classified in the previous tasks. For example, if a sample is incorrectly classified as being *Informative* in Task 1, it can still be classified as belonging to *Non-Infrastructural Damage* class in Task 2, and therefore *Little or No Damage* in Task 3.

Task	Technique	Precision	Recall	F1-Score	AUC
Task 1	**Text RCNN**	0.96	0.96	0.96	0.98
	Image Inception-v3	0.82	0.82	0.82	0.87
	Image $\oplus$ Text (Crisis-DIAS-1)	**0.99**	**0.99**	**0.99**	**1.00**
Task 2	**Text RCNN**	0.96	0.97	096	0.98
	Image Inception-v3	0.92	0.92	0.92	0.91
	Image $\oplus$ Text (Crisis-DIAS-2)	**0.99**	**0.99**	**0.99**	**0.99**
Task 3	**Text RCNN**	0.95	0.95	0.95	0.96
	Image Inception-v3	0.95	0.94	0.94	0.95
	Decision Fusion-Equal	0.73	0.73	0.73	0.76
	Decision Fusion-F1	0.95	0.95	0.95	0.97
	Feature Fusion	0.96	0.97	0.97	0.97
	Attention Fusion (Crisis-DIAS-3)	**0.96**	**0.98**	**0.97**	**0.98**

Table 1: Results and ablation study.

5 Results and Discussion

5.1 Experimental Setup

We use Stratified 5 fold cross-validation to establish our results. We also use SMOTE (Chawla et al. 2002) on the word embeddings to handle the class imbalance in the training folds for linguistic baselines. For the RCNN, we use LSTM layer with hidden dimension 64 to capture the contextual dependencies. The final feature vector dimension (before the softmax layer) is 128 in case of text models and 1024 for image models. We train the models using early stopping with a batch size of 64. We use Adam optimizer with an initial learning rate of 0.001, and the values of $\beta 1$ and $\beta 2$ as 0.9 and 0.999, respectively. All the experiments were run on a GeForce GTX 1080 Ti GPU with memory speed of 11 Gbps.

5.2 Results

To demonstrate the effectiveness of Crisis-DIAS for multimodal damage assessment on social media, we perform the following independent studies:

Ablation Study. Table 1 highlights the results of an ablation study over the best linguistic and vision models along with Crisis-DIAS for the three tasks to demonstrate the effectiveness of multimodal damage assessment models. For each task, the *fusion* model forms a constituent of Crisis-DIAS (named as *Crisis-DIAS-*X); these are then combined, utilizing gates to form the final Crisis-DIAS framework, shown in Figure 1.

Design Choices. We tried different architectures for modelling text- CNN (Kim 2014), hierarchical attention model, bidirectional LSTM and RCNN (Lai et al. 2015). However, in the interest of brevity, we discuss the results for the RCNN model, which performed the best on all the three tasks. The architecture considerably reduces the effect of noise in social media posts (Lai et al. 2015). As input to the model, we use 100-dimensional Fasttext word embeddings (Bojanowski et al. 2016) trained on the dataset. By operating at character n-gram level, Fasttext tends to capture the morphological structure well, which helps the otherwise out of vocabulary words (such as *hash-tags*) to share semantically similar embeddings with its component words. For images, InceptionV3 performed the best which employs multiple sized filters to get a *thick* rather than a *deep* architecture, as very deep networks are prone to overfitting. Such a design makes the network computationally less expensive, which is a prime concern for Crisis-DIAS as we want to minimize latency to give quick service to the disaster relief groups.

Fusion Results. For Tasks 1 and 2, the linguistic and vision pipelines are combined using the logical OR operation ($\oplus$). The evaluation in Table 1 clearly shows the considerable enhancement in results attained by combining the two modalities.

For Task 3, we experimented with three fusion techniques, namely- Decision fusion, Feature fusion, and Attention fusion. Although combining different modalities for improving the performance appears to be intuitively appealing, in practice can be challenging (Wang et al. 2007) due to the varying levels of noise and conflicts between different modalities (Atrey et al. 2010). The results of Task 3 in Table 1 summarize the impact of the augmentation made to the image classifier. Weighted maximum decision fusion does not lead to a substantial improvement in the performance of the unimodal system due to the presence of conflicting noise in the two modalities. Both Feature and Attention fusion results in a boost in the evaluation metrics by combining the latent representations of both the modalities. Attention fusion further learns which features are more *important* than the others and has higher recall than the former, which is why we choose Attention fusion for the final framework of Crisis-DIAS.

System Level Analysis. Although, a hierarchical system like Crisis-DIAS is capable of handling the incorrect predictions of a task in the later stages, yet in some cases, these errors may propagate the chain till the end and never be caught. This is one limitation of pipeline structures where the performance of a task depends on all the ones before it.

Therefore, in order to evaluate the whole system, we define the system level F1-score, F_s, as the product of the F1-scores for all the individual tasks involved in the pipeline,

$$F_s = F_1 * F_2 * F_3 \tag{7}$$

where F_1, F_2 and F_3 are the F1-scores of the three tasks, respectively as in Table 1. The formula is justified as it discounts the performance of the current task by the error rate of the entire pipeline before it. F_s for Crisis-DIAS is 0.95.

For a system designed to provide aid in crucial times, the latency for the complete end-to-end prediction must be very *less*. To analyze Crisis-DIAS in this respect, we find the time taken by each task in predicting a random sample of 1000 tweets. Table 2 shows the average time per task and the total average time for 10 iterations. Evidently, on an average, Crisis-DIAS can identify and assess damage in nearly 100 tweets in 1 minute, which is double the speed of the uni-

Task	Average Time (seconds)
Task 1	216
Task 2	158
Task 3	275
Average total time: 649 seconds	

Table 2: Crisis-DIAS time analysis

Model	Precision	Recall	F1-Score
Unimodal Text CNN (Sreenivasulu and Sridevi 2019)	0.76	0.76	0.76
Proposed Unimodal Text RCNN	0.96	0.96	0.96
Proposed Multimodal Crisis-DIAS-1	**0.99**	**0.99**	**0.99**

Table 3: Comparison with previous work on Task 1.

modal vision framework proposed by (Alam, Ofli, and Imran 2018b).

Comparison and Evaluation. The objective of this work is to identify and provide emergency response services to where damage is severe, and therefore, even the slightest error can make the entire task even more *disastrous* for the support teams. Consequently, the objective is to have a small false negative rate. The increase in the values for recall obtained by combining the two pipelines helps in reducing this rate for the three tasks.

We compare our results with those obtained by past research on the dataset. (Sreenivasulu and Sridevi 2019) used the dataset for identifying the informative posts by utilizing only the text (Task 1). They employ CNN and ANN to achieve the best F1 score of 0.76. Both the linguistic pipeline (0.96) and the multimodal technique (0.99) proposed in this work outperform their results (refer Table 3).

6 Discussion

Here we discuss some of the practical aspects related to the deployment of Crisis-DIAS and the limitations that they induce.

Responsiveness. Systems like Crisis-DIAS must not only be effective in correctly identifying and analyzing the damage in social media posts, they must also do it efficiently. From the system-level analysis conducted in the previous section, we observed how our proposed multimodal framework takes considerably less time in comparison to an existing unimodal system, and therefore, is fairly responsive.

Demographic Bias. One of the primary inadequacies of the system, is that it only analyzes the posts suffering from *infrastructural and utility damage* for the severity of impact, whereas other *humanitarian* categories such as *affected individuals* and *vehicle damage* could also reflect the presence of severe damage but have not been considered. The imbalance of social media information about disasters is highly skewed and biased based on geographical and socio-economic factors. A more generic identification pipeline based on GANs for adding new classes and data augmentation paves our future work.

Computational Efficiency. The models for Tasks 1 and 2 classify a tweet as positive if there is the slightest indication for the presence of damage. Although we could have directly analyzed all the samples for damage severity (Task 3), the amount of data generated on social media platforms is huge. Therefore, it would have been computationally very expensive to use an attention network for all the samples, further delaying the process of resource allocation. To maintain a trade-off between efficiency and effectiveness, we use a simple logical OR ($\oplus$) operation to integrate the two modalities. Doing so, we not only can discard several tweets with no characteristic damage present but also ensure that it is done only when both the linguistic and vision classifiers are confident about the prediction. For Task 3, we use attention fusion, slightly more convoluted but efficacious at the same time, to identify the severity of damage with high precision.

Credibility. A damage assessment system that relies on social media content is heavily dependent on the fact that the content is true. This consideration applies to all other systems build to utilize social media data. No user can be held responsible in case the information posted by them is false, which can often be the case. Identifying the truthfulness of the content, however, is another non-trivial task.

Access & Remote Deployment. Lastly, such a system fails to cater to disaster situations which occur in remote areas that are sparsely populated, or where people do not have access to social media.

Generalization for unimodal tweets. The pipeline, as it stands, can also identify the presence of damage if either of the modality is missing *i.e.,* Task 1 and 2 will work seamlessly for unimodal tweets. Since Crisis-DIAS has inherently been trained only on multimodal tweets from the Crisis-MMD dataset, using attention and feature fusion at the Task 3 level for unimodal tweets may output undesirable garbage values. Therefore, Task 3 for such unimodal tweets can be done using Decision Fusion with a weight of 0 assigned to the missing modality.

7 Qualitative Analysis

To further our analysis, we discuss the following examples which help in justifying the design choices made for each of the three tasks involved in the Crisis-DIAS framework. The correctly predicted labels are followed by a check-mark (✓) and the incorrect predictions are followed by a cross (✘) and the expected true label inside parentheses *(actual)*.

- **QA-1:** In this example, both the linguistic and vision classifiers correctly predict it as *informative* for Task 1 and therefore, it is considered for damage analysis in Task

Figure 3: **QA-1: Text**- *"Mora" Leaves A Trail of Destruction Across Teknaf 02-06-2017 https://t.co/1e3NnJqZnW*
Crisis-DIAS Prediction: Informative✓ | Infrastructural damage✓
Image Prediction: Informative✓ | Non-infrastructural damage✘*(Infrastructural damage)*
Text Prediction: Informative✓ | Infrastructural damage✓

2. The image classification model incorrectly classifies it as *not* having *infrastructural damage*, as the image depicts a storm with no other *damage indicating feature (like debris, etc)*. The text, on the other hand, gives a sense of damage by using words like *'destruction*, and is correctly classified to the *infrastructural damage* class. Crisis-DIAS classifies the tweet as a whole as having *infrastructural damage*, as desired.

- **QA-2:**

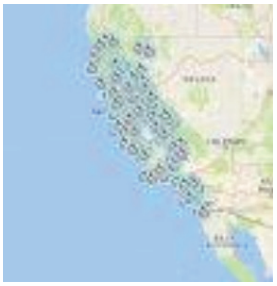

Figure 4: **QA-2:** **Text**- *RT @Rincon_Music: Radio Reports On California Wildfires - https://t.co/JKlnvNczM8 https://t.co/yIRhhwft6z*
Crisis-DIAS Prediction: Informative✘*(Non-informative)* | Infrastructural damage✘*(Non-infrastructural damage)* | No damage✓
Image Prediction: Informative✘*(Non-informative)* | Infrastructural damage✘*(Non-infrastructural damage)* | Mild damage✘*(No damage)*
Text Prediction: Informative✘*(Non-informative)* | Infrastructural damage✘*(Non-infrastructural damage)* | Mild damage✘*(No damage)*

In this tweet, both the text and image classifiers of Tasks 1 and 2 fail to correctly classify the tweet as non-informative and having no infrastructural damage, respectively. Despite that, the sample is correctly classified as having no damage by the attention fusion model of Task 3 due to the hierarchical framework of Crisis-DIAS.

8 Error Analysis

In this section, we continue our analysis by discussing some cases where the proposed Crisis-DIAS architecture is not able to model the desired intricacies and gives inaccurate results in comparison to other unimodal and multimodal techniques, elaborating the plausible reason for the same.

- **EA-1:**

  Due to the incorrect prediction by the image classifier in Task 2, the sample is a false negative for Crisis-DIAS as the image model confuses the example with those originally belonging to the *vehicle damage* class, now labeled as *non-infrastructural damage*.

- **EA-2:**

  The fusion model in Crisis-DIAS Task 3 emphasizes on damage presence-based features in a tweet. Words like *death* and broken walls, are examples of such features which might be the cause of Crisis-DIAS overestimating the severity of the damage.

Figure 5: **EA-1:** **Text**- *President Xi sends condolences over earthquake along Iran-Iraq earthquake https://t.co/alJ1zagxJQ #XiJinping https://t.co/v1L8yt29uF*
Crisis-DIAS Prediction: Informative✓ | Non-infrastructural damage✘*(Infrastructural damage)*
Image Prediction: Informative✓ | Non-infrastructural damage✘*(Infrastructural damage)*
Text Prediction: Informative✘*(Non-informative)* | Non-infrastructural damage✓

Figure 6: **EA-2:** **Text**- *Death toll from #Irma is 34; but still counting. https://t.co/q76ayumni6 https://t.co/0zquN1hr6T*
Crisis-DIAS Prediction: Informative✓| Infrastructural damage✓| Severe damage✘*(Mild damage)*
Image Prediction: Informative✓| Infrastructural damage✓| Mild damage✓
Text Prediction: Informative✓| Infrastructural damage✓| Mild damage✓

9 Conclusion

In this work, we propose a gated multimodal deep learning framework, Crisis-DIAS, for identifying and analyzing the level of damage severity in social media posts with the scope for betterment in disaster management and planning. We characterize Crisis-DIAS along the following dimensions: **Multimodal System-** The proposed system leverages from both textual and visual cues to automate the process of damage identification and assessment from social media data. **Efficient and Effective-** We perform an extensive analysis of the proposed system over the CrisisMMD dataset (Alam, Ofli, and Imran 2018a) by showing how fusing the features from different media sources helps the proposed multimodal framework to correctly identify the samples which were otherwise missed by the unimodal classifiers. **Responsive and Real-time-** Based on the system analysis, the proposed framework is not only better in terms of its performance, but

is also faster and can, therefore, be deployed for real-time assessment. **Generic-** The system is generic and can be used for multiple disaster scenarios.

References

Ahmad, K.; Pogorelov, K.; Riegler, M.; Conci, N.; and Halvorsen, P. 2019. Social media and satellites. *Multimedia Tools and Applications* 78(3):2837–2875.

Alam, F.; Ofli, F.; Imran, M.; and Aupetit, M. 2018. A twitter tale of three hurricanes: Harvey, irma, and maria. *ArXiv* abs/1805.05144.

Alam, F.; Ofli, F.; and Imran, M. 2018a. Crisismmd: Multimodal twitter datasets from natural disasters. *CoRR* abs/1805.00713.

Alam, F.; Ofli, F.; and Imran, M. 2018b. Processing social media images by combining human and machine computing during crises.

Alam, F.; Ofli, F.; and Imran, M. 2019. Crisisdps: Crisis data processing services. *Proceedings of the 16th ISCRAM Conference*.

Asvadi, A.; Garrote, L.; Premebida, C.; Peixoto, P.; and Nunes, U. J. C. 2017. Multimodal vehicle detection: fusing 3d-lidar and color camera data. *Pattern Recognition Letters* 115:20–29.

Atrey, P. K.; Hossain, M. A.; El Saddik, A.; and Kankanhalli, M. S. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems* 16(6):345–379.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *CoRR* abs/1607.04606.

Chaudhuri, N., and Bose, I. 2019. Application of image analytics for disaster response in smart cities. 3036–3045.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16(1):321–357.

Chew, C., and Eysenbach, G. 2010. Pandemics in the age of twitter: Content analysis of tweets during the 2009 h1n1 outbreak. *PLoS ONE 5(11): e14118.*

Conover, M. D.; Davis, C.; Ferrara, E.; McKelvey, K.; Menczer, F.; and Flammini, A. 2013. The geospatial characteristics of a social movement communication network.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Global Digital Report. 2019. Global digital report 2019.

Hua, X.-S., and Zhang, H.-J. 2004. An attention-based decision fusion scheme for multimedia information retrieval. In *Pacific-Rim Conference on Multimedia*, 1001–1010. Springer.

Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; and Vieweg, S. 2014. Aidr: artificial intelligence for disaster response. In *WWW*.

Imran, M.; Castillo, C.; Diaz, F.; and Vieweg, S. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)* 47(4):67.

Kim, Y. 2014. Convolutional neural networks for sentence classification. *CoRR* abs/1408.5882.

Lai, S.; Xu, L.; Liu, K.; and Zhao, J. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 2267–2273. AAAI Press.

Llorente, A.; Garcia-Herranz, M.; Cebrian, M.; and Moro, E. 2015. Social media fingerprints of unemployment. *PLoS ONE 10(5): e0128692.*

Mouzannar, H.; Rizk, Y.; and Awad, M. 2018. Damage identification in social media posts using multimodal deep learning. *Proceedings of the 15th ISCRAM Conference*.

Olteanu, A.; Castillo, C.; Diaz, F.; and Vieweg, S. 2014. Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.

Pouyanfar, S.; Tao, Y.; Tian, H.; Chen, S.-C.; and Shyu, M.-L. 2019. Multimodal deep learning based on multiple correspondence analysis for disaster management. *World Wide Web* 22(5):1893–1911.

Reuter, C., and Kaufhold, M.-A. 2018. Fifteen years of social media in emergencies: A retrospective review and future directions for crisis informatics. *Journal of Contingencies and Crisis Management* 26(1):41–57.

Rizk, Y.; Jomaa, H. S.; Awad, M.; and Castillo, C. 2019. A computationally efficient multi-modal classification approach of disaster-related twitter images. In *SAC*.

Said, N.; Ahmad, K.; Regular, M.; Pogorelov, K.; Hasan, L.; Ahmad, N.; and Conci, N. 2019. Natural disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications* 1 – 36.

Sreenivasulu, M., and Sridevi, M. 2019. Detecting informative tweets during disaster using deep neural networks. *International Conference on Communication Systems & Networks*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Vempala, A., and Preoţiuc-Pietro, D. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2830–2840. Florence, Italy: Association for Computational Linguistics.

Wang, S.; Dash, M.; Chia, L.-T.; and Xu, M. 2007. Efficient sampling of training set in large and noisy multimedia data. *ACM Trans. Multimedia Comput. Commun. Appl.* 3(3).

Wang, X.; Gong, G.; and Li, N. 2017. Multimodal fusion of eeg and fmri for epilepsy detection. *IJMSSC* 9:1850010.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, 3320–3328.