AI for Disaster Rapid Damage Assessment from Microblogs

Muhammad Imran,[1] Umair Qazi,[1] Ferda Ofli,[1] Steve Peterson,[2] Firoj Alam[1]

[1]Qatar Computing Research Institute, Doha, Qatar
[2]Montgomery County Community Emergency Response Team, Maryland, USA
mimran@hbku.edu.qa, uqazi@hbku.edu.qa, fofli@hbku.edu.qa, stevepeterson2@gmail.com, fialam@hbku.edu.qa

Abstract

Formal response organizations perform rapid damage assessments after natural and human-induced disasters to measure the extent of damage to infrastructures such as roads, bridges, and buildings. This time-critical task, when performed using traditional approaches such as experts surveying the disaster areas, poses serious challenges and delays response. This paper presents an AI-based system that leverages citizen science to collect damage images reported on social media and perform rapid damage assessment in real-time. Several image processing models in the system tackle non-trivial challenges posed by social media as a data source, such as high-volume of redundant and irrelevant content. The system determines the severity of damage using a state-of-the-art computer vision model. Together with a response organization in the US, we deployed the system to identify damage reports during a major real-world disaster. We observe that almost 42% of the images are unique, 28% relevant, and more importantly, only 10% of them contain either mild or severe damage. Experts from our partner organization provided feedback on the system's mistakes, which we used to perform additional experiments to retrain the models. Consequently, the retrained models based on expert feedback on the target domain data helped us achieve significant performance improvements.

Introduction

At disaster onset, humanitarian organizations seek to assess the impacts of the disaster. One crucial task that they perform is rapid damage assessment—preferably in the first 72 hours. The rapid damage assessment task is a prerequisite of several response operations (FEMA 2021) and helps first responders understand affected areas for planning immediate rescue and relief operations. Traditional damage assessment methods require field assessments by experts who locate damaged infrastructure, interview people, and collect other relevant data. These experts perform analysis and interpretation of the gathered data before writing a report for planners and decision-makers. However, this process is usually challenged by limited human resources and severe conditions in the disaster area. These challenges disrupt data collection, analysis, damage assessment, and consequently, delay relief operations.

The use of technology for rapid damage assessment includes remote sensing through synthetic aperture radar or optical imagery (Plank 2014; Barrington et al. 2012; Pesaresi, Gerhardinger, and Haag 2007). However, these costly data sources are time-consuming to deploy and collect relevant data. Furthermore, satellite data is susceptible to noise such as clouds, especially during weather-induced disasters like hurricanes. This work employs non-traditional data sources such as social networks to acquire citizen-generated data during disasters in real-time to address the challenges mentioned above. More specifically, as opposed to using textual content for damage detection (Kryvasheyeu et al. 2016), we utilize imagery content shared during disaster events to identify scenes that show damages.

Studies show that images shared on Twitter carry information pertinent to damage detection and severity assessment for humanitarian response (Alam, Ofli, and Imran 2018a). Therefore, this paper presents an AI-based system, called Rapid Damage Assessment (RDA)[1], for real-time analysis of images shared on Twitter during disasters. Specifically, the system uses state-of-the-art computer vision models to perform several image processing tasks. These include image de-duplication, relevant image identification, and damage severity assessment. To test the effectiveness and performance of the damage severity assessment model, we deployed the system during a real-world disaster with an emergency response team. Specifically, we activated the RDA system during 2019 Hurricane Dorian in collaboration with Montgomery County, Maryland Community Emergency Response Team (MCCERT). The activation focused on identifying damage severity at three levels: (i) severe damage, (ii) mild damage, and (iii) little-to-no damage (i.e., none).

Domain experts from our partner organization examined the output of the system and provided two types of feedback. First, they verified the system's output and reported if a prediction was correct or not. Second, in case of an incorrect prediction, experts provided the correct label. Experts feedback, i.e., the correct labels for images where the system made mistakes, was used to retrain the damage assessment model. We performed several experiments to show the performance of our existing model when applied to the experts' annotated test set. Moreover, new models are trained

[1]https://rda-aidr.qcri.org/hurricane_dorian_2019.php

with and without data augmentation strategies to demonstrate their performance degradation due to shift in domain distributions. The model trained to incorporate images annotated by domain experts outperforms all other models, which highlights the need and importance of the target domain data. More details of the deployment in terms of different kinds of mistakes that the experts identified, challenges that the models faced, and potential directions to address those challenges are described in (Imran et al. 2020).

Community Emergency Response Team

Community Emergency Response Teams (CERTs) offer a consistent, nationwide approach to volunteer training that professional responders can rely on during disaster situations (Gov. 2021). CERTs assist formal humanitarian organizations in a range of disaster response and management tasks. For example, CERTs expand their team capabilities to provide virtual assistance that includes social media analysis. Montgomery County, Maryland CERT applies a methodological framework as described in (Peterson et al. 2019) when searching for mission-specific content extracted from Twitter. This includes, but is not limited to, the following tasks to find reports of damage:

1. *Use hashtags and keywords to manually search for relevant tweets, including tweets containing images showing some degree of damage.*
2. *Analyze tweet text for pertinent cues that would qualify it as valuable (e.g., context, location, user profile, etc.).*
3. *Download damage images into a team collaborative working document and determine the applicability of each image to the mission assignment.*
4. *Send summary-of-findings report, including appropriate images, to the respective stakeholder (e.g., FEMA).*
5. *Repeat above steps throughout operational period.*

The above-described methodological framework is effective for social media analysis during disasters, when the mission assignment is focused on text. For example, searching tweets for information indicating road conditions within a disaster-hit region. Most social media management tools that Montgomery County, Maryland CERT has used lack the capability to retrieve only tweets containing disaster images. This hinders mission assignments related to retrieving visual data because of complex and time-consuming manual steps. For example, first, each tweet would need to be individually checked by a human to determine if an image was included. Second, if the tweet did contain an image, and that image was determined to be of value to the mission assignment, it would need to be extracted and placed within a collaborative document. Then, another human analyzes the kind of impact shown in the image and determines the applicability to the mission assignment.

Manual analysis of a high-volume data source such as Twitter often leads to information overload (Hiltz and Plotnick 2013). Therefore, instead of following the above manual steps, we used an automatic Twitter image collection and processing system to find reports of damages caused by Hurricane Dorian as it was progressing. Next, we describe the details of the automatic processing system.

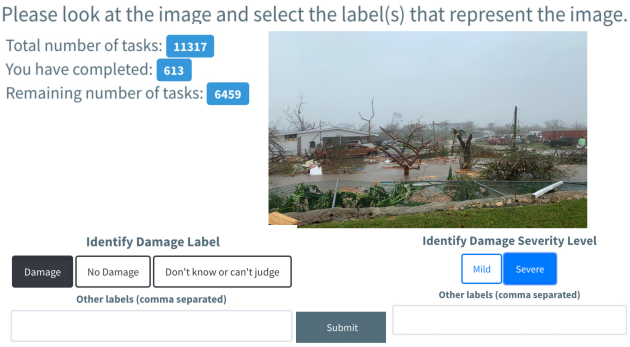

Figure 1: Annotation interface for both tasks

AI-based Damage Assessment System

Analyzing social media image streams in real-time is challenging due to overwhelming noise, redundant content, and a high data rate. We built the Rapid Damage Assessment (RDA) system with several computational components to deal with the aforementioned challenges. RDA is a major extension of our existing social media monitoring system, called AIDR (Imran et al. 2014b). Next, we describe different image processing components of the system.

RDA Image Processing Components

The RDA system relies on five components, including one data collection and four data processing components.

Data Collector The Twitter streaming API is used to collect real-time tweets during a disaster situation. More than one data collector can be initiated to capture multiple disaster streams simultaneously. Tweet collection match either keywords/hashtags, geographical bounding boxes, or posts from specific users—defined per disaster event basis.

Image URL Deduplicator Retweets and re-sharing on Twitter produce redundant image URLs resulting in high download time and potentially requiring ample storage. The image URL deduplicator maintains a hash of unique URLs through which duplicate links are detected. Specifically, when a new image URL arrives, the system queries the hash to determine whether it is a unique URL or not. This hash-based search time-complexity is O(1), i.e., the search takes constant time irrespective of the hash queue length. Finally, images corresponding to the unique URLs are downloaded.

Image Deduplicator Images downloaded from unique URLs may not be actually unique. Different URLs pointing to the same image, or an image that is cropped, resized, or re-shared with additional text inserted are some potential causes of image-level duplication. Therefore, performing image-level de-duplication by comparing it with existing images is crucial. The image deduplicator module performs this check by measuring the distance between a newly collected image and all existing images using the Euclidean distance on deep features extracted from images. The system uses a deep neural network to extract image features and keeps them in a hash. We use a fine-tuned VGG16

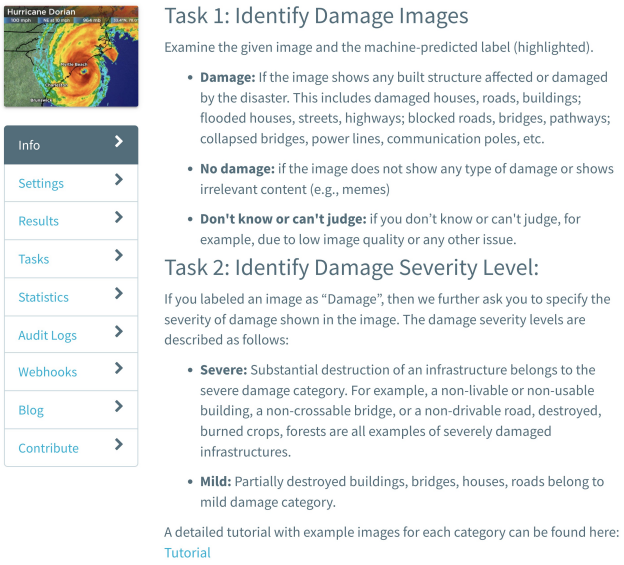

Figure 2: Task description page with class definitions

model (Simonyan and Zisserman 2014) and extract features from its penultimate fully-connected (i.e., "fc2") layer. A Euclidean distance less than 20 between the features of two images is considered as the two images are duplicate or near-duplicate. Determining an optimal distance threshold is an empirical question, which is not the focus of this work. However, a distance of 20 worked best for our setting.

Junk Filtering Large quantities of noisy content, even during disaster events, make it to Twitter. These irrelevant and noisy images usually contain cartoons, advertisements, celebrities, and explicit content as trending hashtags are often exploited for this purpose (Alam, Ofli, and Imran 2018a,b). Disaster decision-makers during response and recovery efforts have limited time. Therefore, identifying and filtering irrelevant content from the system's output is essential. The junk filtering module detects irrelevant images using a deep learning model trained to detect irrelevant concepts such as cartoons, celebrities, banners, and advertisements. The F1-score (i.e., the harmonic mean of the precision and recall) of this model is 98% (Nguyen et al. 2017).

Damage Severity Assessment Finally, images that are unique and relevant are processed by the damage severity assessment module, which determines the level of damage. For this purpose, we fine-tune an existing VGG16 model pre-trained on the ImageNet dataset. The fine-tuning of the network is based on the damage-related labeled dataset with three classes. The distribution of images across the classes is: *severe=11,510*, *mild=3,762*, and *none=10,548*. The *severe damage* class contains images that show fully destroyed houses, buildings, bridges, etc. The *mild damage* class contains images that show partially destroyed scenes of houses, buildings, or transportation infrastructure. The model is tested on a held-out test set (20%), and its performance in terms of macro-precision, macro-recall, and macro-F1 is 0.757, 0.728, and 0.737, respectively[2].

Human-in-the-loop Automatic systems are not perfect and make mistakes. It is essential to have human involvement either to verify the produced results or provide supervision to the system if/when needed (Imran et al. 2014a). Our system uses human-in-the-loop for both verification and supervision purposes. Data items processed by the system are used to take samples for humans to verify and guide the system if a mistake is identified. Such mistakes could be false positives or false negatives. Human-labeled items would then be ideally fed back to the system for retraining a new model for enhanced performance. To involve humans in the verification and supervision process, we use our MicroMappers crowdsourcing system (Lucas et al. 2014).

RDA Deployment for Hurricane Dorian

The RDA system was activated on August 30, 2019 when Hurricane Dorian was a Category-2 storm barreling toward the northern Bahaman Islands and central Florida. In the next 24 hours, the tropical storm rapidly intensified and became a potential danger. On September 1, it made landfall in the Bahamas in Elbow Cay. On September 2, the hurricane remained nearly stationary over the Bahamas as a Category-5 storm. On September 3, the hurricane began weakening as it started moving northwestward, parallel to the east coast of Florida. The hurricane turned to the northeast the next day and made landfall on Cape Hatteras with a Category-1 intensity on September 6. The keywords and hashtags used for data collection included *"HurricaneDorian, Dorian, DorianAlert, PuertoRico, DorianMissing, DorianDeaths, Dorian Found, DorianFound"*, among others.

Next, we set up our MicroMappers platform to involve humans in the verification and supervision processes. Images classified by the system were sampled in batches. Preferably, images with severe and mild damage should be sampled periodically due to two reasons. First, experts' feedback on images from the positive classes is more valuable for understanding the model's weaknesses for the types of images that matter to response organizations. Second, periodic samples will help us get feedback on a diverse set of images rather than those from a particular day or hour. The sampling strategy ran every couple of hours during the operational period for Montgomery County, Maryland CERT. For most samples, *severe damage* and *mild damage* images were selected, however, when they were not available we also sampled images from *none* class.

Images in a varying time window of past N-hours formed a sample set. We did not fix the number of hours, N, as human processing speed depended on many unknown factors. The experts then examined sampled images for one of the three damage levels described above. Images along with system predictions were shown on a web interface for experts' evaluation. The experts either agreed or disagreed with machine predictions. In case of disagreement, experts were required to provide the correct label, as well. Figure 1 illustrates the annotation interface, which shows *Damage*,

[2]The weighted F1-score of the model is 0.83

Date	Images collected	Unique	Relevant	Mild damage	Severe damage
Aug 30, 2019	15,255	7,347 (48.16%)	3,591 (23.54%)	681 (4.46%)	295 (1.93%)
Aug 31, 2019	27,064	9,272 (34.26%)	5,819 (21.50%)	1,281 (4.73%)	549 (2.03%)
Sep 1, 2019	26,612	8,135 (30.57%)	5,804 (21.81%)	1,432 (5.38%)	587 (2.21%)
Sep 2, 2019	43,337	13,859 (31.98%)	9,936 (22.93%)	2,073 (4.78%)	870 (2.01%)
Sep 3, 2019	32,757	10,597 (32.35%)	8,661 (26.44%)	1,768 (5.4%)	1,052 (3.21%)
Sep 4, 2019	29,312	9,662 (32.96%)	8,371 (28.56%)	1,515 (5.17%)	1,033 (3.52%)
Sep 5, 2019	33,630	18,054 (53.68%)	10,018 (29.79%)	2,506 (7.45%)	1,354 (4.03%)
Sep 6, 2019	24,545	14,557 (59.31%)	7,864 (32.04%)	1,705 (6.95%)	1,484 (6.05%)
Sep 7, 2019	14,030	8,820 (62.87%)	4,685 (33.39%)	813 (5.79%)	881 (6.28%)
Sep 8, 2019	8,900	5,229 (58.75%)	3,768 (42.34%)	545 (6.12%)	704 (7.91%)
Sep 9, 2019	7,709	4,653 (60.36%)	3,068 (39.80%)	384 (4.98%)	706 (9.16%)
Sep 10, 2019	5,666	3,289 (58.05%)	2,110 (37.24%)	210 (3.71%)	439 (7.75%)
Sep 11, 2019	3,923	2,203 (56.16%)	1,379 (35.15%)	139 (3.54%)	334 (8.51%)
Sep 12, 2019	3,424	1,940 (56.66%)	1,233 (36.01%)	151 (4.41%)	441 (12.88%)
Sep 13, 2019	2,929	1,829 (62.44%)	974 (33.25%)	116 (3.96%)	246 (8.40%)
Sep 14, 2019	726	321 (44.21%)	299 (41.18%)	23 (3.17%)	69 (9.50%)
Total	**279,819**	**119,767 (42.8%)**	**77,580 (27.73%)**	**15,342 (5.48%)**	**11,044 (3.95%)**

Table 1: Daily distribution of images collected and identified as unique, relevant, and with mild or severe damage.

No Damage, and a *Don't know or can't judge* options. If an expert selects the *Damage* label, the interface asks to select one of two severity levels (*Mild*, *Severe*), shown on the right side of the screen. Experts were allowed to provide additional comments using a text box on the interface.

In addition to the labeling interface, we established two other pages, one for showing the task details (Figure 2) and another for a detailed tutorial[3] with concrete examples for each class. Each human expert was instructed to go through the tutorial before labeling.

Deployment Results

The joint activation of the RDA system started on August 30, 2019 and ran for almost two weeks. The system collected 6,890,106 tweets, out of which 280,063 unique image URLs were obtained. The total number of downloaded images was 279,819. Around 244 images failed to download for various reasons, such as the tweet author deleted the actual tweet, the image host server was down, or the connection timed out.

Automatic Image Classification Results

RDA is a real-time image processing system, which processes items as they are collected from Twitter. Images downloaded from unique URLs are immediately fed to the image deduplicator to ensure the newly captured images are not duplicate of some previously captured images. The junk filtering component then checks unique images. Finally, the damage severity assessment component ingests unique and potentially relevant images where damage severity is determined. Table 1 presents the daily distribution of collected images and those which were identified as unique, relevant, and containing mild or severe damage.

Throughout the activation period, the system captured and analyzed 279,819 images. Out of which, 119,767 (42%) were found as unique images by the image-based deduplication module. Due to the high retweet/re-sharing ratio on Twitter, 58% of the images were identified as exact or near-duplicate by the system even during a large-scale natural disaster. At this stage, automatically filtering out duplicate images has already substantially reduced the chance of information overload affecting human experts.

Figure 3: Severe damage images found by the RDA system

Furthermore, out of the 279,819 images, 77,580 (27%) were identified as relevant by the system. These images do not contain cartoons, celebrities, banners, advertisements, etc. Among the relevant images, some contained damage scenes while others did not.

Around 26,386 (10%) images were identified as containing some damage where 11,044 (4%) contained severe and 15,342 (6%) mild damage. Filtering out ∼90% of the images as potentially not containing any damage content by the system is a significant reduction in the risk of information overload for humans. Figure 3 shows example images with severe damage while Figure 4 with mild damage.

[3]https://ibb.co/DztXbTy

Figure 4: Mild damage images found by the RDA system

N=28,050		Machine	
		Damage	***No Damage***
Human	***Damage***	7.44%	2.54%
	No Damage	21.23%	68.79%

Table 2: Damage detection task confusion matrix—system vs. human judgments

N=28,050		Machine		
		Severe	***Mild***	***None***
Human	***Severe***	2.53%	1.37%	1.27%
	Mild	0.40%	3.14%	1.27%
	None	2.57%	18.66%	68.79%

Table 3: Damage severity assessment task confusion matrix—system vs. human judgments

Human Verification and Image Labeling Results

In total, 28 experts from the response organization examined the evolving samples taken from the system processed image stream over 42 hours from 8pm on September 6 to 2pm on September 8. Since our annotators are trained emergency managers, we trusted their judgments without asking multiple assessors. However, at the end of the operational period, the team lead of the experts reviewed about 2,000 of the completed tasks for quality assurance.

Table 2 reports results of the first task (i.e., damage detection). The experts verified and provided their feedback for 29,136 images. Recall, these images were initially processed by the system and contained scenes of both damage and no damage. Moreover, an image with damage content is labeled with one of three damage severity levels (severe, mild, none). In total, 1,086 images were considered "Don't know or can't judge" by the experts due to several reasons, including blurred/low-quality images, close-up shots, too dark/small, or an image containing text. Out of the remaining images (i.e., 28,050), the experts agreed with the system predictions for 21,384 images. We show the details of experts' agreement and disagreement with the system in Table 2. We observed that in 2,088 (7.44%) cases, both system and human agreed that the images show damage, and for 19,296 (68.79%) cases, no damage is visible. Nonetheless, the experts did not agree with the system predictions for 6,666 images (∼25%). The RDA system yields an accuracy of 76% based on the experts' analysis.

For the damage severity task that determines level of damage severity in an image, we show results in Table 3. In terms of experts' agreement with the system, we observed that for 20,887 images, experts agreed with the system predictions. However, there are 7,163 images where experts disagreed with the system prediction for a particular severity level. Based on the human analysis, the RDA system yields an accuracy of 74% in this task.

Retraining Models Using Expert Annotations

Prior studies report model inefficiencies when applied in the wild. Several reasons could potentially harm models' performance, including differences in the data distributions of source and target domains. Although the damage severity assessment model used in this activation was trained on social media images from past disasters, quantifying its performance in different evaluation metrics (e.g., F1-score) is essential to understand potential weaknesses in the target domain (i.e., Hurricane Dorian). Moreover, as the annotations obtained from experts constitute a precious dataset that strictly follows the definition of damage and its severity levels, we sought to determine performance gains by retraining our existing model with the expert annotations.

To this end, the expert annotations (i.e., 28,050 images) are divided into train, development, and test sets with 70%, 10%, and 20% ratios, respectively. Table 5 shows the distribution of expert annotations into three sets. Next, we perform four types of experiments. Since our existing model was trained without using any data augmentation strategies, we designed two new experiments that ignore data augmentation to make the obtained results directly comparable with the existing results. However, two additional experiments are designed to use data augmentation strategies as they tend to positively impact models' performance.

Table 4 reports results from four models trained using different training sets. We used a combination of old (Old_{train}) and new ($HDorian_{train}$) data, and trained models with and without data augmentation. We tested all four models on the new test set ($HDorian_{test}$). Not surprisingly, Model-1, which is trained on old train set without data augmentation, yields low performance (i.e., F1=0.570) when applied on the new Hurricane Dorian test set. Next, Model-2 is trained using the old training set but with data augmentation. Strangely, the data augmentation did not help, and consequently, the model's performance dropped (i.e., F1=0.547).

To determine whether the new training set from Hurricane Dorian deployment has any positive impact on model performance, next we train two new models using both old and new training sets. Model-3 is trained without data augmentation whereas Model-4 uses data augmentation. Again, both models are tested on the same Hurricane Dorian test set. Even though no data augmentation was used for Model-3, it clearly shows substantial improvement over Model-1 and Model-2. Furthermore, Model-4 outperforms all other mod-

Models	Data Aug.	Training set	Test set	Macro-Prec.	Macro-Rec.	Macro-F1
Model 1	No	Old_{train}	$HDorian_{test}$	0.526	0.684	0.570
Model 2	Yes	Old_{train}	$HDorian_{test}$	0.510	0.650	0.547
Model 3	No	$Old_{train} + HDorian_{train}$	$HDorian_{test}$	0.669	0.691	0.678
Model 4	Yes	$Old_{train} + HDorian_{train}$	$HDorian_{test}$	**0.716**	**0.721**	**0.718**

Table 4: Results obtained from four models trained with and without Hurricane Dorian training set (i.e., $HDorian_{train}$) and data augmentation. All four models are tested on Hurricane Dorian test set (i.e., $HDorian_{test}$).

Class	Train	Dev	Test	Total
Severe	1,016	148	287	1,451
Mild	944	138	267	1,349
None	17,675	2,575	5,000	25,250
Total	19,635	2,861	5,554	28,050

Table 5: Data splits of Hurricane Dorian annotated images.

els and achieves a plausible macro F1-score of 0.718. These experiments evidently show that the training data from the target event helps minimize discrepancies between source and target domain distributions. Moreover, closing the loop by using expert feedback to retrain the system is important from the emergency managers' point of view to use a better model during future disasters.

Related Work

The importance of imagery content for disaster response has been reported in a number of studies (Turker and San 2004; Chen et al. 2013; Plank 2014; Feng et al. 2014; Fernandez Galarreta, Kerle, and Gerke 2015; Attari et al. 2017; Erdelj and Natalizio 2016; Ofli et al. 2016). These studies dominantly analyze aerial and satellite imagery data. For instance, (Turker and San 2004) analyze post-earthquake aerial images to detect damaged infrastructure caused by the August 1999 Izmit earthquake in Turkey. Another study provides a comprehensive overview of multi-temporal Synthetic Aperture Radar procedures for damage assessment and highlights the advantages of SAR compared to the optical sensors (Plank 2014).

On the other hand, there are studies that report the importance of images captured by Unmanned Aerial Vehicles (UAV) for damage assessment while highlighting the limitations of remote sensing data (Fernandez Galarreta, Kerle, and Gerke 2015; Attari et al. 2017). These studies propose per-building damage scores by analyzing multi-perspective, overlapping and high-resolution oblique images obtained from UAVs. In (Ofli et al. 2016), the authors also highlight the importance of UAV images while addressing the limitations of satellite images, and propose a methodology that enables volunteers to annotate aerial images, which is then combined with machine learning classifiers to tag images with damage categories.

Very recently, the study of social media image analysis for disaster response has received attention from the research community (Daly and Thom 2016; Mouzannar, Rizk, and Awad 2018; Alam, Ofli, and Imran 2018b). For example, researchers analyze images extracted from social media data collected during a fire event (Daly and Thom 2016). Specifically, they analyze spatio-temporal meta-data associated with the images and suggest that geo-tagged information is useful to locate the fire-affected areas. Another study investigates damage detection by focusing on human and environmental damages (Mouzannar, Rizk, and Awad 2018). The study includes collecting multimodal social media posts and labeling them with six categories such as (1) infrastructural damage (e.g., damaged buildings, wrecked cars, and destroyed bridges) (2) damage to natural landscape (e.g., landslides, avalanches, and falling trees) (3) fires (e.g., wildfires and building fires) (4) floods (e.g., city, urban and rural) (5) human injuries and deaths, and (6) no damage.

While many of the past works on rapid damage assessment need expensive data sources, some of which are also time consuming to deploy such as UAVs, satellites, our work highlights the usefulness of Twitter images and utilizes an image processing pipeline proposed in (Nguyen et al. 2017). The RDA system filters irrelevant content, removes duplicates, and assesses damage severity for real-time damage assessment using deep learning techniques.

Conclusions

Information about the impacts, particularly damages, caused by a disaster event is essential for response organizations' timely actions. Rapid damage assessment is a crucial task that formal response organizations perform through field assessments and remote sensing methods. Research studies demonstrate that citizen-reported data as text messages and images on social networking platforms contain valuable information about situational awareness and disaster impacts. This work presented RDA, a system to perform rapid damage assessment from Twitter image streams. Together with a response organization, we deployed RDA during a real-world disaster (i.e., Hurricane Dorian) where human experts examined the system's output and provided feedback. In addition to reporting the system's performance during the activation, we performed several experiments to quantify the strengths and weaknesses of our damage assessment model on a test set formed from experts labels. Moreover, we trained several damage assessment models using a combination of old and new training images and data augmentation variations. The results obtained from the model that was trained on combined training data with augmentation outperformed all existing models. This highlights the importance of target domain data for training more robust models.

References

Alam, F.; Ofli, F.; and Imran, M. 2018a. CrisisMMD: Multimodal twitter datasets from natural disasters. In *Proc. of the 12th ICWSM, 2018*, 465–473. AAAI press.

Alam, F.; Ofli, F.; and Imran, M. 2018b. Processing Social Media Images by Combining Human and Machine Computing during Crises. *International Journal of Human–Computer Interaction*, 34(4): 311–327.

Attari, N.; Ofli, F.; Awad, M.; Lucas, J.; and Chawla, S. 2017. Nazr-CNN: Fine-Grained Classification of UAV Imagery for Damage Assessment. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 50–59.

Barrington, L.; Ghosh, S.; Greene, M.; Har-Noy, S.; Berger, J.; Gill, S.; Lin, A. Y.-M.; and Huyck, C. 2012. Crowdsourcing earthquake damage assessment using remote sensing imagery. *Annals of Geophysics*, 54(6).

Chen, T.; Lu, D.; Kan, M.-Y.; and Cui, P. 2013. Understanding and classifying image tweets. In *ACM International Conference on Multimedia*, 781–784. ACM.

Daly, S.; and Thom, J. 2016. Mining and Classifying Image Posts on Social Media to Analyse Fires. In *Proc. of the 13th ISCRAM*, 1–14. ISCRAM Association.

Erdelj, M.; and Natalizio, E. 2016. UAV-assisted disaster management: Applications and open issues. In *2016 international conference on computing, networking and communications (ICNC)*, 1–5. IEEE.

FEMA. 2021. Preliminary Damage Assessment Guide. https://www.fema.gov/media-library/assets/documents/109040. Accessed: 2021-11-04.

Feng, T.; Hong, Z.; Fu, Q.; Ma, S.; Jie, X.; Wu, H.; Jiang, C.; and Tong, X. 2014. Application and prospect of a high-resolution remote sensing and geo-information system in estimating earthquake casualties. *Natural Hazards and Earth System Sciences*, 14(8): 2165–2178.

Fernandez Galarreta, J.; Kerle, N.; and Gerke, M. 2015. UAV-based urban structural damage assessment using object-based image analysis and semantic reasoning. *Natural Hazards and Earth System Sciences*, 15(6): 1087–1101.

Gov., U. 2021. Community Emergency Response Team. https://www.ready.gov/cert. Accessed: 2021-11-04.

Hiltz, S. R.; and Plotnick, L. 2013. Dealing with information overload when using social media for emergency management: Emerging solutions. In *Proc. of the 10th ISCRAM, 2013*. ISCRAM.

Imran, M.; Alam, F.; Qazi, U.; Peterson, S.; and Ofli, F. 2020. Rapid damage assessment using social media images by combining human and machine intelligence. In *Proceedings of the 17th International Conference on Information systems for crisis response and management (ISCRAM)*.

Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; and Rogstadius, J. 2014a. Coordinating human and machine intelligence to classify microblog communications in crises. In *ISCRAM*.

Imran, M.; Castillo, C.; Lucas, J.; Meier, P.; and Vieweg, S. 2014b. AIDR: Artificial intelligence for disaster response. In *Proc. of the ACM Conference on WWW*, 159–162. ACM.

Kryvasheyeu, Y.; Chen, H.; Obradovich, N.; Moro, E.; Van Hentenryck, P.; Fowler, J.; and Cebrian, M. 2016. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3): e1500779.

Lucas, J.; Imran, M.; Castillo, C.; and Meier, P. 2014. MicroMappers. https://micromappers.qcri.org/. Accessed: 2021-11-04.

Mouzannar, H.; Rizk, Y.; and Awad, M. 2018. Damage Identification in Social Media Posts using Multimodal Deep Learning. *15th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2018)*, (May): 529–543.

Nguyen, D. T.; Alam, F.; Ofli, F.; and Imran, M. 2017. Automatic Image Filtering on Social Networks Using Deep Learning and Perceptual Hashing During Crises. In *International Conference on Information Systems for Crisis Response and Management (ISCRAM)*.

Ofli, F.; Meier, P.; Imran, M.; Castillo, C.; Tuia, D.; Rey, N.; Briant, J.; Millet, P.; Reinhard, F.; Parkan, M.; et al. 2016. Combining human computing and machine learning to make sense of big (aerial) data for disaster response. *Big data*, 4(1): 47–59.

Pesaresi, M.; Gerhardinger, A.; and Haag, F. 2007. Rapid damage assessment of built-up structures using VHR satellite data in tsunami-affected areas. *International Journal of Remote Sensing*, 28(13-14): 3013–3036.

Peterson, S.; Stephens, K.; Hughes, A.; and Purohit, H. 2019. When Official Systems Overload: A Framework for Finding Social Media Calls for Help during Evacuations. In *Proceedings of the Information Systems for Crisis Response and Management Conference*, 867–875.

Plank, S. 2014. Rapid damage assessment by means of multi-temporal SAR—A comprehensive review and outlook to Sentinel-1. *Remote Sensing*, 6(6): 4870–4906.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Turker, M.; and San, B. T. 2004. Detection of collapsed buildings caused by the 1999 Izmit, Turkey earthquake through digital analysis of post-event aerial photographs. *International Journal of Remote Sensing*, 25(21): 4701–4714.