Rapid Damage Assessment Using Social Media Images by Combining Human and Machine Intelligence

Muhammad Imran*
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
mimran@hbku.edu.qa

Firoj Alam
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
fialam@hbku.edu.qa

Umair Qazi
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
uqazi@hbku.edu.qa

Steve Peterson
Montgomery County, Maryland Community
Emergency Response Team
United States
stevepeterson2@gmail.com

Ferda Ofli
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
fofli@hbku.edu.qa

ABSTRACT

Rapid damage assessment is one of the core tasks that response organizations perform at the onset of a disaster to understand the scale of damage to infrastructures such as roads, bridges, and buildings. This work analyzes the usefulness of social media imagery content to perform rapid damage assessment during a real-world disaster. An automatic image processing system, which was activated in collaboration with a volunteer response organization, processed ~280K images to understand the extent of damage caused by the disaster. The system achieved an accuracy of 76% computed based on the feedback received from the domain experts who analyzed ~29K system-processed images during the disaster. An extensive error analysis reveals several insights and challenges faced by the system, which are vital for the research community to advance this line of research.

Keywords

Social Media, Damage Assessment, Artificial Intelligence, Image Processing.

INTRODUCTION

Rapid damage assessment is a task that humanitarian organizations perform within the first 48 to 72 hours of a disaster and is considered a prerequisite of many disaster management operations[1]. Assessing the severity of damage helps first responders understand affected areas and the extent of impact for the purpose of immediate rescue and relief operations. Moreover, based on the results of early damage assessment, humanitarian organizations identify focus areas to make detailed assessment for long-term relief and rehabilitation of the affected population[2]. However,

*Corresponding author
[1] https://www.fema.gov/media-library/assets/documents/109040
[2] http://www.resiliencenw.org/2012files/LongTermRecovery/DisasterAssessmentWorkshop.pdf

traditional ways to perform rapid damage assessment require sending experts to the disaster affected area to conduct field assessments that include taking pictures of the damaged infrastructure, interviewing people, and collecting relevant data from other reliable sources. These experts perform analysis and interpretation of the gathered data before writing a report for planners and decision-makers. Limited human resources and severe living conditions in the disaster area are only a few examples of challenges facing field assessment experts. Such challenges can delay data gathering, damage assessment, and ultimately, relief operations.

Past works on rapid damage assessment used Synthetic Aperture Radar (SAR), remote sensing, and satellite image processing techniques (Plank 2014; Barrington et al. 2012; Pesaresi et al. 2007). These approaches use costly data sources and are time consuming to deploy and collect relevant data. Furthermore, satellite data is susceptible to noise such as clouds, especially during weather-induced disasters, e.g., hurricanes. The focus of this work is to analyze the usefulness of non-traditional data sources such as social media to perform rapid damage assessment. More specifically, as opposed to textual content for damage detection (Kryvasheyeu et al. 2016), we are interested in the imagery content shared during an ongoing natural disaster to identify images that contain damage caused by the disaster.

Microblogging and social media platforms such as Twitter play an increasingly important role during disasters (Castillo 2016; Imran, Castillo, Diaz, et al. 2015). People turn to Twitter to get updates about an ongoing emergency event (Starbird et al. 2010; Hughes and Palen 2009). More importantly, when people in disaster areas share information about what they witness in terms of damages caused by the disaster, flooded streets, reports of missing, trapped, injured or deceased people, or other urgent needs, that information could potentially be leveraged by humanitarian organizations to gain situational awareness and to plan relief operations (Imran, Elbassuoni, et al. 2013; Purohit et al. 2014).

In addition to the textual messages, images shared on Twitter carry important information pertinent to humanitarian response. This work focused on the real-time analysis of the imagery data shared on Twitter during Hurricane Dorian in 2019. In collaboration with a volunteer response organization, Montgomery County, Maryland Community Emergency Response Team (MCCERT)[3], we activated our image processing system before Hurricane Dorian made landfall in the Bahamas. Based on the information requirements of our partner organization, the system filtered images that were relevant to the disaster and identified the ones that showed some damage content (e.g., damaged buildings, roads, bridges). More specifically, the damage analysis task assessed the severity of the damage using three levels: (i) severe damage, (ii) mild damage, and (iii) little-to-no damage (i.e., none).

During a 13-day deployment period, the system collected around ~280K images. It used machine learning techniques to eliminate duplicate and irrelevant images before performing the damage assessment. As a result, around ~160K images were found as relevant and around ~26K as containing some damage content. Domain experts from our partner volunteer response organization examined an evolving sample of images during the disaster. The purpose of having human-in-the-loop was two-fold. *First*, to keep an eye on the system generated output to verify the system was correctly classifying the images and make corrections if a mistake was identified. *Second*, use the human corrections to better train the system for future deployments.

The human experts performed two tasks while examining over ~29K images over several days during the system's deployment period. *First*, they determined if an image contained any damage content. *Second*, if an image was identified as containing damage, they would determine the severity of the damage using the three severity levels mentioned above.

Based on the results of each expert's assessment, the system achieved an accuracy of 76% for the damage detection task and 74% for the damage severity assessment task. These are reasonable accuracy scores, which prove the effectiveness of the system for analyzing real-world disaster imagery data for rapid damage assessment. Furthermore, we performed an error analysis of the corrections resulting from performing the two tasks. Among common mistakes, we observed that the system is weak in identifying scenes that show flooding taken from afar, foggy or blurry scenes, and low-light scenes. Moreover, images that resembled damage scenes but were verified as incorrect confused the system. For example, a pile of trash would sometimes be confused as damage. Identifying deficiencies during our deployment not only helps us improve our machine learning models, but also provides valuable information for the crisis informatics research community to better understand challenges of analyzing social media imagery data during real-world disaster situations. This could lead to the discovery of additional methods and models that seek similar qualifying actionable machine output imagery to benefit decision-makers.

The rest of the paper is organized as follows. The next section summarizes *Related Work*. In the *Hurricane Dorian Deployment* section we provide details of the event and our system deployment. Then, we report the data collection

[3]We met the lead of the CERT team in one of the ISCRAM conferences and discussed the possibility to do a joint activation of our automatic image processing system for damage assessment.

and analysis in the section *Data and Results*. We later discuss our findings in the *Discussion* section, identify challenges, and provide future directions. Finally, we conclude the paper in the last section.

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 et al. 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. Plank 2014 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.

On the other hand, Fernandez Galarreta et al. 2015 and Attari et al. 2017 report the importance of images captured by Unmanned Aerial Vehicles (UAV) for damage assessment while highlighting the limitations of remote sensing data. These studies propose per-building damage scores by analyzing multi-perspective, overlapping and high-resolution oblique images obtained from UAVs. Ofli et al. 2016 also highlights the importance of UAV images while addressing the limitations of satellite images. The authors 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 et al. 2018; Alam et al. 2018b). For example, Daly and Thom 2016 analyze images extracted from social media data collected during a fire event. 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. Mouzannar et al. 2018 investigate damage detection by focusing on human and environmental damages. Their 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, and SAR, our work highlights the usefulness of Twitter images and utilizes an image processing pipeline proposed in (Nguyen et al. 2017). This image processing system filters irrelevant content, removes duplicates, and assesses damage severity for real-time damage assessment using deep learning techniques and human-in-the-loop.

HURRICANE DORIAN DEPLOYMENT

Hurricane Dorian

On the morning of August 30, 2019, Hurricane Dorian was a Category 2 in the eastern Caribbean 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, Dorian began weakening in intensity 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. It then transitioned into an extra-tropical cyclone and struck Nova Scotia and then Newfoundland with hurricane-force winds on September 8. Finally it dissipated near Greenland on September 10. Hurricane Dorian caused 63 direct and 7 indirect fatalities. It caused USD 8.28 billion worth of damage. Affected areas included Lesser Antilles, Puerto Rico, The Bahamas, Eastern United States, Eastern Canada, southern Greenland, and Iceland[4].

Community Emergency Response Team Deployment

In the United States, Community Emergency Response Teams (CERTs) offer a consistent, nationwide approach to volunteer training and organization that professional responders can rely on during disaster situations[5]. When called upon, CERTs can assist formal humanitarian organizations in a range of disaster response and management tasks. Some CERTs have expanded their team capabilities to provide virtual assistance that includes social media analysis. Montgomery County, Maryland CERT applies a methodological framework as described by (Peterson et al. 2019) when searching for mission-specific content extracted from Twitter. This includes, but is not limited to, performing the following tasks to find reports of damage:

[4] https://en.wikipedia.org/wiki/Hurricane_Dorian
[5] https://www.ready.gov/cert

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 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 region impacted by the disaster. Most social media management tools that Montgomery County, Maryland CERT has used lack the capability to retrieve only tweets containing disaster images. This could hinder future 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. Secondly, 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 potentially another human would determine the applicability of the image 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.

Automatic Image Processing System Deployment

We used AIDR image processing system (Nguyen et al. 2017; Imran, Castillo, Lucas, Meier, and Vieweg 2014) to start collecting tweets related to Hurricane Dorian on August 30, 2019. The collection ran for about two weeks and stopped on September 14, 2019. In total, approximately 6,890,106 tweets were collected. The below listed keywords were used to collect English language tweets.

Keywords used to collect tweets
HurricaneDorian, Dorian, DorianAlert, Alerts_Dorian, PuertoRico, DorianMissing, DorianDeaths, HurricaneDorianMissing, HurricaneDorianDeaths, Dorian Missing, Hurricane Dorian Missing, Dorian Found, DorianFound

Images Processing Modules

The system has a number of different image analysis modules to process images on Twitter. In this system deployment, we used four of them, which are described below.

***Image URL deduplication*:** Due to the high number of retweets, some images are re-shared on Twitter thousands of times. Downloading duplicate or near-duplicate images is time-consuming and not helpful for decision-makers. This module keeps track of image URLs and maintains a hash of unique ones. Upon receiving a new image URL, the system determines whether it is unique or not by querying the hash with a time-complexity of O(1)[6] i.e., the search takes constant time irrespective of the hash queue length.

***Image deduplication*:** Images downloaded using unique URLs are not warranted to be actually unique. Different URLs can point to the same image hosted and shared by different web hosts. Moreover, an image could be cropped, resized, or re-shared with additional text inserted on the existing image. Therefore, determining whether an image is a duplicate by comparing it to all the existing images collected by the system to date is crucial. The image deduplication module performs this check by measuring the distance between a newly collected image and existing images using the Euclidean distance on features extracted from images. More specifically, the system uses a deep neural network to extract features from an image and keeps it in a hash. We use a fine-tuned VGG16 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

[6] https://en.wikipedia.org/wiki/Time_complexity

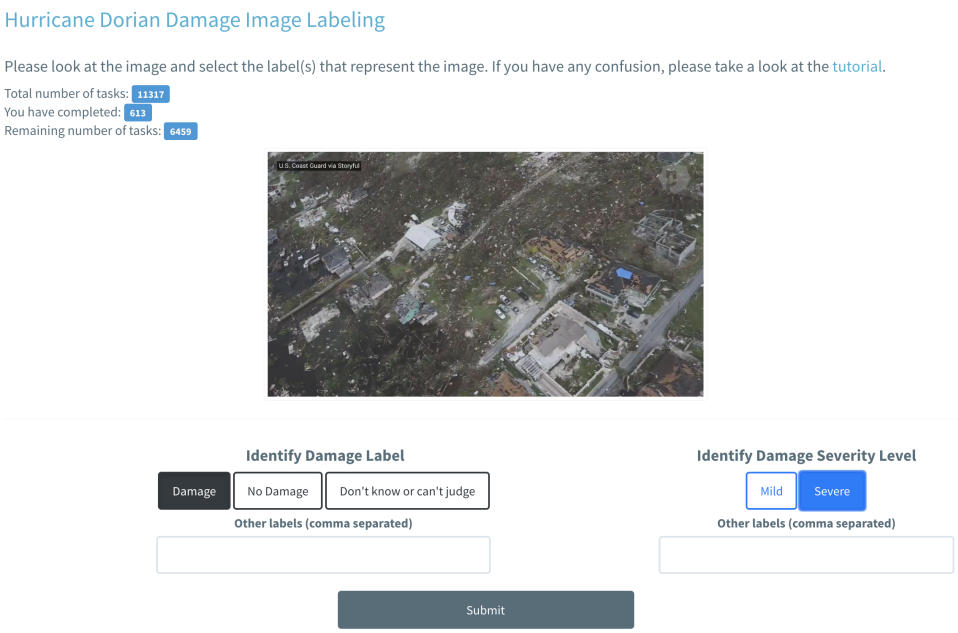

Figure 1. Web-based interface for human assessors to verify system predictions and relabel images if required. The highlighted labels are system predicted.

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*:** Generally, Twitter is full of noisy content and disasters are not an exception. Research studies have found images of cartoons, advertisements, celebrities, and explicit content shared in tweets related to a disaster event (Alam et al. 2018a; Alam et al. 2018b). Trending hashtags are often exploited for this purpose. Such irrelevant content must not be shown to decision-makers during disaster response and recovery efforts given their time is valuable and limited. Unnecessary disruptions must be avoided. The junk filtering module tries to detect irrelevant images by using a deep learning model which is trained to detect irrelevant concepts such as cartoon, celebrities, banners, 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*:** A unique and potentially relevant image is then finally analyzed by the damage severity assessment module, which determines the level of damage shown in the image. We used a transfer learning technique to fine-tune an existing VGG16 model originally trained on the ImageNet dataset. The fine-tuning of the network (all layers) is performed based on the damage-related labeled dataset consisting of three classes. The three classes are *severe*, *mild*, and *none*. The *severe damage* class contains images that show fully destroyed houses, building, bridges, etc. The *mild damage* class contains images that show partially destroyed scenes of houses, building, or transportation infrastructure. The F1 score of this model is 83%.

Human-in-the-loop for Image Labeling in Real Time

Automatic systems are not perfect and may make mistakes. It is essential to have some human involvement either to verify the produced results or provide supervision to the system if/when needed (Imran, Castillo, Lucas, Meier, and Rogstadius 2014). Our system uses human-in-the-loop for both verification and gaining 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 used our MicroMappers crowdsourcing system[7]. Images downloaded and classified by our data processing system were first used to take samples. We performed this sampling every couple of hours during the operational period for Montgomery County, Maryland CERT (details in the next section). In most of the samples, we selected all *severe damage* and *mild damage* images and some from the *none* class from the system-processed images in a given time-window of past T-hours. We did not fix

[7] https://micromappers.qcri.org/

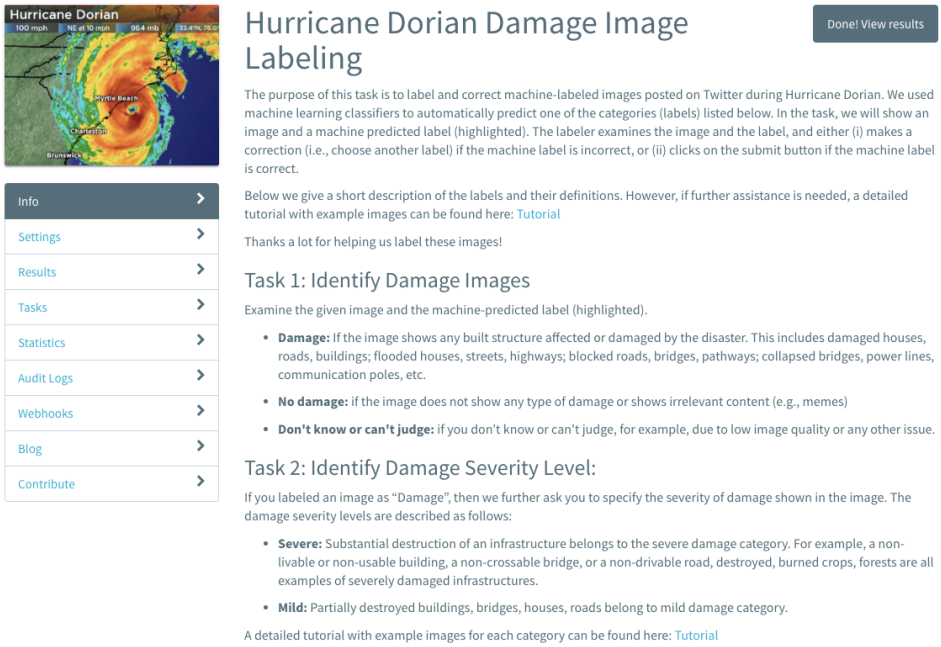

Figure 2. Task description page showing details of the tasks including classes definition

the T value, i.e., number of hours, as human processing speed depends on many unknown factors. The sampled images were then shown to human experts. For this deployment, we decided to only crowdsource the output of our damage severity assessment module, which classifies an image into one of three damage levels (i.e., classes), as described above. On a web interface, we showed an image along with the system predicted class to the expert. The human expert either agreed or disagreed with the machine classification. In the case where they disagreed with the machine classification, they would provide a new label to the image. Figure 1 depicts the crowdsourcing interface. The interface first showed the options (*Damage*, *No Damage*, and *Don't know or can't judge*), which can be seen on the left. If the human selected the *Damage* label, the interface would further show two severity levels (*Mild*, *Severe*), which would appear on the right side of the screen. The human would select one of these two severity labels and submit their assessment. If the human selected *"Don't know or can't judge"*, then the system would not show the two additional severity labels. The human experts were allowed to provide additional comments using the text boxes on the interface.

In addition to the labeling interface, we established two other pages, one for showing the task details (Figure 2) and the second for a detailed tutorial[8] with concrete examples for each class. Each human expert was instructed to go through the tutorial before beginning their labeling effort.

DATA AND RESULTS

Data Statistics

As shown in Table 1, out of all 6,890,106 tweets collected, 280,063 unique image URLs were found. The total number of downloaded images was 279,819. Around 244 images failed to download due to one of several reasons, e.g., the tweet author deleted the actual tweet, the image host server was down, or connection timed out, etc.

Table 1. Hurricane Dorian tweets and image data statistics

Total tweets	Unique image URLs	Downloaded images	Failed to download
6,890,106	280,063	279,819	244

Automatic Classification Results

The 279,819 images, which were successfully downloaded, were then analyzed by the image processing modules described in the previous section. An image-based deduplication was performed as the first step followed by the

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

image relevancy check executed by the *Junk filtering* module. All relevant images were then ingested by the *Damage Severity Assessment* module to determine if they contained any damage. If the image contained damage, it was classified according to the severity shown in the scene.

Table 2. Image-based automatic processing: Results of unique, relevant, and damage images.

Unique images	Relevant images	Images with damage	Severe damage	Mild damage
119,767	77,580	26,386	11,044	15,342

Table 2 shows the number of images which were found as unique, relevant, and with some level of damage – specifically severe and mild damage. Out of 279,819 images, the image-based deduplication module found 119,767 unique images, which was around 42% of the whole set. As described earlier, this image-based deduplication module relies on deep features extracted from images using a deep neural network. Due to the high retweet/re-sharing ratio on Twitter, even during a large-scale natural disaster, 58% of the images were identified as exact or near-duplicate by the system. At this stage, the process of automatically finding near-duplicate images had already reduced the chance of information overload affecting the human experts.

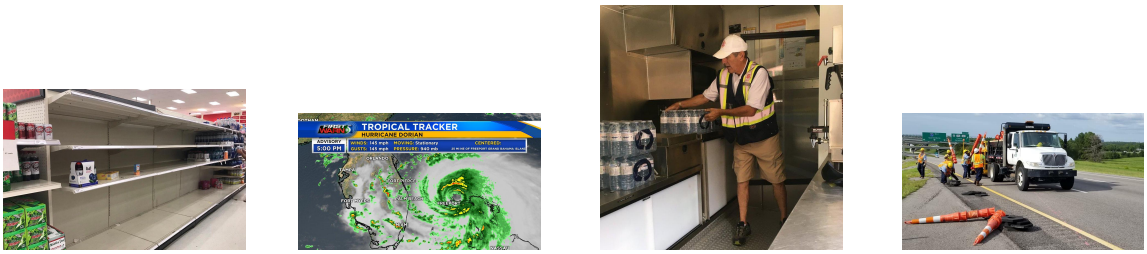

Figure 3. Images that are relevant but do not show any damage

Furthermore, out of the 279,819 images, 77,580 were identified as relevant by the system. These images did not contain cartoons, celebrities, banners, advertisements, etc. Among the relevant images, some contained damage scenes while others did not. Figure 3 shows a few images that did not show any damage but were identified as relevant. Many of the relevant images showed hurricane maps or some other scene associated to rescue efforts. We show the distribution of total, relevant, and irrelevant images for the whole deployment period in Figure 4.

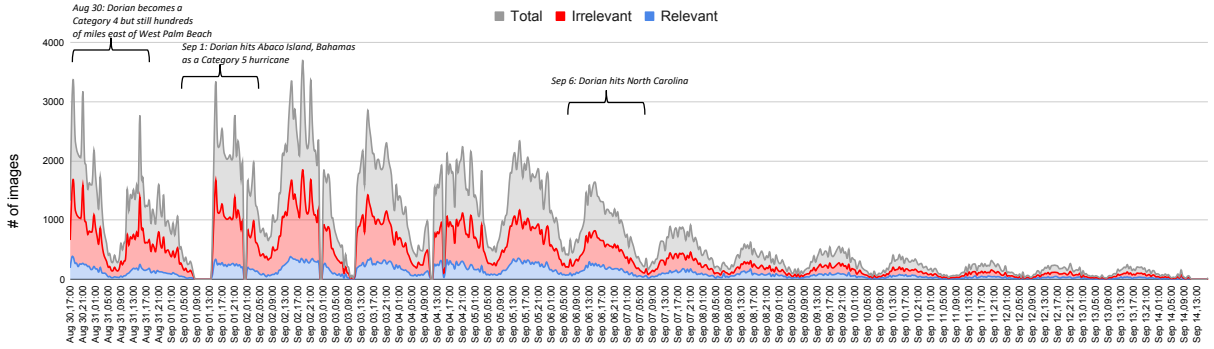

Figure 4. Distribution of total, relevant, and irrelevant images over the duration of the event

Out of all relevant images, 26,386 were identified as containing some damage where 11,044 showed severe and 15,342 showed mild damage. The images with damage scenes were around ~10% of all the downloaded images. The system's ability to filter out ~90% of images as potentially not containing any damage content is a significant reduction in risk of information overload to humans. Figure 5 contains a few images, which according to the system showed severe damage. Figure 6 shows a few images, which according to the system included mild damage. We show the distribution of mild and severe damage images as classified by the system for the whole duration period in Figure 7.

Finally, Table 3 shows distribution of images identified as duplicate, not relevant, and containing no damage.

Figure 5. Severe damage images identified by the automatic system

Figure 6. Mild damage images identified by the automatic system

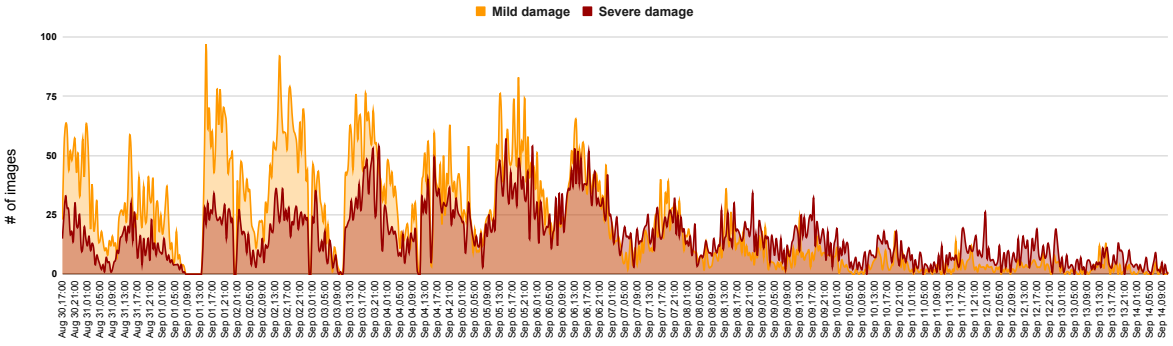

Figure 7. Distribution of severe and mild damage images over the duration of the event

Table 3. Image-based automatic processing: Results of duplicate, not relevant, and no damage images

Duplicate images	Not relevant images	Images with no damage
160,052	202,239	253,433

Human Verification and Image Labeling Results

As Hurricane Dorian progressed, human experts from Montgomery County, Maryland CERT (N=28) were asked to look at an evolving sample of system processed images to verify if the system was producing desired results. The experts were also instructed to correct any identified mistakes done by the system. Given they were trained domain experts, not employed from an online paid crowdsourcing platform, we trusted their judgements without the need to ask multiple assessors. This meant each image was assessed by only one human expert. At the conclusion of the CERT operational period, their team lead reviewed around ~2K of the completed tasks for quality assurance. This feedback can be found in the *Discussion* section.

Table 4. Damage detection task confusion matrix—system vs. human judgments

N=28,050		**Machine**	
		Damage	***No Damage***
Human	***Damage***	2,088	712
	No Damage	5,954	19,296

Table 4 shows the results of the damage detection task. In total, the human experts analyzed 29,136 images over a 42-hour operational period from 8:00pm on September 6 to 2:00pm on September 8. These images were initially processed by the system and contained scenes of both damage and no damage. Moreover, when an image contained damage, it had one of three damage severity labels (severe, mild, none) assigned by the system. Of all 29,136 analyzed images, 1,086 were labeled as "Don't know or can't judge" by the experts. This could have been due to several reasons including blurred/low quality images, closeup shots, too dark/small, or an image containing text. From the remaining set (i.e., 28,050), the experts agreed with the system predictions for 21,384 images. This agreement can be seen in the two diagonal colored cells of the Table 4, where in 2,088 cases both system and human agreed that the image showed some damage and 19,296 cases the image showed no damage. However, there were 6,666 (5,954 + 712) images which the experts did not agree with the system. Based on the results of this human analysis, we compute the system accuracy = 76% .

Table 5. Damage severity assessment task confusion matrix—system vs. human judgments

N=28,050		**Machine**		
		Severe Damage	***Mild Damage***	***None***
Human	***Severe Damage***	710	384	357
	Mild Damage	113	881	355
	None	721	5,233	19,296

For the second task, which aimed to assess the severity of damage in an image, the results are shown in Table 5. The human experts agreed with the system 20,887 times, as shown in the three diagonal colored cells. However, we received a disagreement for 7,163 images. Based on the results of this human analysis, we measured the system accuracy as 74% for this task.

We report detailed system performance results in terms of precision, recall, F1, and accuracy for both tasks in Table 6. The system achieved a precision of 0.89 for both tasks, which is a reasonable score. However, the recall scores are a little lower, i.e., 0.76 for task 1, and 0.74 for task 2.

DISCUSSION AND ERROR ANALYSIS

From an emergency manager's point of view, it is important the system does not miss any damage reports, regardless of the severity of damage. Missed damage reports could provide relevant information on an impacted area that had

Table 6. System performance for both tasks

Classification tasks	Accuracy	Precision	Recall	F1
Task 1: Damage; No damage	0.76	0.89	0.76	0.80
Task 2: Severe; Mild; None	0.74	0.89	0.74	0.80

minimal or no actionable intelligence immediately available for decision-making. Therefore, among other cases, 'false negative' are most important for us to analyze. For example, when the machine predicts an image as not containing any damage i.e., *"None"* but the human expert labels it as *"Severe"* or *"Mild"*. There were 357 cases where Machine=None & Human=Severe and 355 cases where Machine=None & Human=Mild. These cases can be seen in Table 5 and are analyzed next.

Machine:None vs. Human:Severe: Figure 8 shows a few images where the machine prediction was *None* (i.e., no damage) and human assessment was *Severe damage*. Our in-depth analysis of these 357 images reveals that in most of these false-negative cases, the machine mainly missed flooded scenes. Another main pattern that emerged is where images with low light confused the machine such as the third image from the left in Figure 8. Also, aerial images covering a wide area caused issues for the machine to understand them (i.e., first image on the left). Image collages are also a source of problem for the machine. We define an image collage as multiple images joined together to appear as one. Such cases create even more challenges to accurately classify damage severity when the level of damage in at least one of the images contradicts the level of damage in another image within the same collage.

Figure 8. False negative examples where Machine:None & Human:Severe

Machine:None vs. Human:Mild: Figure 9 shows a few cases out of 355 where the machine prediction was *None*, but according to the human experts these images showed *Mild damage*. Our analysis of these cases revealed that most of the damage appearing in an image was covered by another object. This caused issues for the machine to classify them as a damage image. In the second and third image from the left in Figure 9, it can be seen there are people standing and covering some parts of damage scenes. Whereas in the fourth image, a white door is covering 80% of the damage scene, leaving only a small area for the machine to predict it as a mild damage case. Moreover, we noticed that scenes with trees showing strong winds were also missed by the machine.

For the above two cases, further investigation revealed that our damage severity assessment model's training data lacks flooded and strong winds scenes, which is one of the reasons the model missed many such cases.

Figure 9. False negative examples where Machine:None & Human:Mild

While responding to a disaster, time is limited and precious to decision-makers. Having them preoccupied with looking at irrelevant data is not appropriate and risks creating information overload consequences. Therefore, another important area for us to understand are false positives our system generates. As shown in Table 5, there were 5,954 (721 + 5,233) images which, according to the machine, either contained *Severe damage* (i.e., 721 cases)

or *Mild damage* (i.e., 5,233 cases), but according to the human experts these cases were *None*—meaning they did not show any damage. Next, we study these two cases.

Machine:Severe vs. Human:None: We extensively analyzed these 721 images. A few of them are shown in Figure 10. Our analysis revealed that most of the images appeared to contain some damage, but actually they did not. Many images contained scenes with irregular arrangements of wooden pieces, which deceived the model to predict them as damage. The first image from the left in Figure 10 shows a pile of trash that could be interpreted as debris of a destroyed built infrastructure. The second image has a wooden pathway with irregular arrangements of lumber. Perhaps, part of the pathway is slightly damaged, but it is not a severe damage scene. Similarly, the other two images caused confusion for the model. Having an ability to identify non-damage scenes which resemble damage scenes would be one of the most challenging tasks to address from the machine's modeling point of view. More hard negative examples would help models better understand and discriminate between positive and negative cases.

A pile of trash that looks like debris of destroyed buildings | A wooden pathway | A dog on small wood pieces or bark | Small wood pieces

Figure 10. False positive examples where Machine:Severe & Human:None

Machine:Mild vs. Human:None: Figure 11 shows a few images from this category. Our analysis revealed that the majority of these images showed maps depicting the hurricane's path as seen in the first image from the left of Figure 11. Among other scenes, there were rough sea images or memes with flooding scenes (i.e., last image from left). Furthermore, we also noticed this category contained many images with people standing in groups or performing some activity. Overall, this category shows more variation in the scenes compared to the other categories. The misclassifications relating to the maps and images where there are people can be easily fixed by feeding more hard negative examples to the machine. However, scenes of rough seas and flooding closely border between the mild or severe categories and thus would be hard to accurately tackle.

Hurricane path map | Sand/soil and white bags | Rough sea | Meme

Figure 11. False positive examples where Machine:Mild & Human:None

Challenges and Future Work

Based on feedback from the human experts, we identified a number of weaknesses and challenges that our image processing models faced. We list these challenges as future work below.

• **Flood scene variations:** Capturing different variations in flood scenes such as flooding on roads, in houses, forests, or fields is important yet challenging for machine learning models. In our case, this problem occurred mainly due to the lack of appropriate training data that represented such variations. However, in some cases, even a sufficient amount of labeled data might not be enough to resolve ambiguities between a natural scene and a disaster scene. For example, rough sea scenes should not be confused as flood scenes. These difficult cases require

additional considerations while training machine learning models and also raise awareness to the need for further research on effective integration of human intelligence into machine learning models.

• **Low-light damage scenes:** We noticed many foggy and low-light scenes were missed by our models. Similar to the previous challenge, lack of training data collected in low-light conditions caused our models to miss such cases. Addressing this issue is important from a time perspective for decision-makers when a disaster occurs at night time. Accurately classified images can provide awareness of the severity of damage before daylight arrives, thus saving time and allowing for some decisions to begin to be made (e.g., resource allocation planning). In addition to collecting more appropriate training data, other image processing techniques can be used to adjust image contrast, brightness, or saturation as pre-processing steps before feeding them into the model.

• **Wide-area and aerial images:** Images taken from afar often cover a wide area that shows many objects such as houses, trees, sky, etc. These images do not only show objects at a much smaller scale than ground-level images but they also often contain scenes with a mix of both damaged and undamaged objects and areas. Due to such large differences in the scale of objects and areas, it may not be ideal to design a single model that operates on both aerial and ground-level images for any given task. In particular for the damage assessment task, the ideal solution may require designing separate aerial and ground-level models with more localized (i.e., object-level) damage detection and assessment capabilities.

• **Maps and memes:** Our models suffered while identifying maps and memes. However, we noticed this deficiency is mainly due to the lack of appropriate labeled data on which our models were initially trained. Adding more suitable training images would help eradicate this problem.

• **Damage-resembling scenes:** Images that show scenes resembling damaged objects or areas constitutes a big challenge for automatic image processing models. We identified around 700 such cases during our deployment. Machine learning for such scenes may need additional semantic information about objects surrounding a damage scene to help models understand. For example, if a nearby crop field shows intact healthy crops, then it is less likely that overall image shows severe damage.

CONCLUSIONS

Rapid damage assessment provides crucial information about damage severity caused by a disaster in the early stages of response. Humanitarian and formal response organizations rely on field assessment reports, remote sensing methods, or satellite imagery to perform damage assessment. This work leveraged imagery data shared on Twitter to identify reports of damages using image processing techniques based on deep neural networks. Moreover, the image processing system filters out duplicate and irrelevant images which are not useful for decision-makers responding to the disaster. The system was activated before Hurricane Dorian made landfall in the Bahamas and ran for 13 days. Over a 42-hour operational period of collaborating with our partner volunteer response organization, the damage reports identified by the system were examined by the domain experts of this organization, whose feedback revealed that the system achieved an accuracy of 74% and 76% for the two damage assessment tasks. Although these scores show the system's effectiveness to process real-world disaster data, we identified a number of shortcomings of our machine learning models, which are listed in the previous section and considered as potential future work.

REFERENCES

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

Alam, F., Ofli, F., and Imran, M. (2018b). "Processing Social Media Images by Combining Human and Machine Computing during Crises". In: *International Journal of Human–Computer Interaction* 34.4, pp. 311–327. eprint: https://doi.org/10.1080/10447318.2018.1427831.

Attari, N., Ofli, F., Awad, M., Lucas, J., and Chawla, S. (Oct. 2017). "Nazr-CNN: Fine-Grained Classification of UAV Imagery for Damage Assessment". In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 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". In: *Annals of Geophysics* 54.6.

Castillo, C. (2016). *Big Crisis Data*. Cambridge, UK: Cambridge University Press.

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

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

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

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". In: *Natural Hazards and Earth System Sciences* 14.8, pp. 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". In: *Natural Hazards and Earth System Sciences* 15.6, pp. 1087–1101.

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.

Hughes, A. L. and Palen, L. (2009). "Twitter adoption and use in mass convergence and emergency events". In: *International Journal of Emergency Management* 6.3-4, pp. 248–260.

Imran, M., Castillo, C., Diaz, F., and Vieweg, S. (2015). "Processing social media messages in mass emergency: A survey". In: *ACM Computing Surveys* 47.4, p. 67.

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

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

Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media". In: *Proc. of the 12th ISCRAM*.

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". In: *Science advances* 2.3, e1500779.

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

Nguyen, D. T., Alam, F., Ofli, F., and Imran, M. (May 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". In: *Big data* 4.1, pp. 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". In: *International Journal of Remote Sensing* 28.13-14, pp. 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*, pp. 867–875.

Plank, S. (2014). "Rapid damage assessment by means of multi-temporal SAR—A comprehensive review and outlook to Sentinel-1". In: *Remote Sensing* 6.6, pp. 4870–4906.

Purohit, H., Hampton, A., Bhatt, S., Shalin, V. L., Sheth, A. P., and Flach, J. M. (2014). "Identifying seekers and suppliers in social media communities to support crisis coordination". In: *Computer Supported Cooperative Work (CSCW)* 23.4-6, pp. 513–545.

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

Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. (2010). "Chatter on the red: what hazards threat reveals about the social life of microblogged information". In: *ACM Conference on Computer Supported Cooperative Work*, pp. 241–250.

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". In: *International Journal of Remote Sensing* 25.21, pp. 4701–4714.