Progress in Disaster Science 2 (2019) 100030

Contents lists available at ScienceDirect

Progress in Disaster Science

journal homepage: www.elsevier.com/locate/pdisas

Using a combination of human insights and 'deep learning' for real-time disaster communication

Brett W. Robertson [a,*], Matthew Johnson [b], Dhiraj Murthy [c], William Roth Smith [a], Keri K. Stephens [a]

[a] Department of Communication Studies, Moody College of Communication, The University of Texas at Austin, 2504A Whitis Avenue, #A1105, Austin, TX 78712, USA
[b] Department of Electrical and Computer Engineering, Cockrell School of Engineering, The University of Texas at Austin, USA
[c] Computational Media Lab, School of Journalism, Moody College of Communication, and Department of Sociology, The University of Texas at Austin, USA

ARTICLE INFO

Article history:
Received 27 January 2019
Received in revised form 26 April 2019
Accepted 4 May 2019
Available online 22 May 2019

Keywords:
Social media
Disasters
Deep learning
Content analysis
Twitter
Images
Rescue

ABSTRACT

Using social media during natural disasters has become commonplace globally. In the U.S., public social media platforms are often a go-to because people believe: the 9-1-1 system becomes overloaded during emergencies and that first responders will see their posts. While social media requests may help save lives, these posts are difficult to find because there is more noise on public social media than clear signals of who needs help. This study compares human-coded images posted during 2017's Hurricane Harvey to machine-learned 'deep learning' classification methods. Our framework for feature extraction uses the VGG-16 convolutional neural network/multilayer perceptron classifiers for classifying the urgency and time period for a given image. We find that our qualitative results showcase that unique disaster experiences are not always captured through machine-learned methods. These methods work together to parse through the high levels of non-relevant content on social media to find relevant content and requests.

1. Introduction

Hurricane Harvey, which struck the greater Houston, Texas, USA-region between August and September 2017, was considered the first *post-911* American natural disaster where social media posts requesting aid and help superseded 9-1-1 phone systems [34]. This form of help-seeking behavior on public social media platforms is over a decade old, but in this particular disaster, social media proved to be a particularly visible, often image-heavy way for stakeholders to disseminate information quickly [41]. In fact, during natural disasters in general, the images shared on social media serve as a type of near-real-time social sensor [26]. Text generally has much higher levels of 'noise', i.e., non-relevant information, than do images posted during a disaster as most images taken during a hurricane, for example, can be categorized under a finite set of motifs [26], whereas text posted during the same time tends to have far more diversity.

During Hurricane Harvey, requesting help on social media proved effective, as various platforms provided citizens with up-to-the second information [17]. For example, hashtags "#SOSHarvey" and "#HelpHouston" trended nationwide during the storm, and were often used as a means to flag those who needed rescue [34]. It is of course no surprise that social media emerged as a visible platform to request assistance, as the American public uses social media in the course of their everyday lives [31]. Following changing perceptions towards the visibility of content on social media and the turn towards online surveillance cultures, an American Red Cross study found that people believe that emergency personnel monitor social media and that their calls for help will be answered if they simply post a message [2]. It is also clear that people, regardless whether they were directly affected by the hurricane, engaged with various social media platforms, and shared content on their own accord as a means of activism, support, or keeping up with other's real-time status [17].

Further research explained that citizens use social media to request aid and relief during a natural disaster [33]. However, Varga et al. [47] stresses that often Twitter messages directed to provide assistance are not successful in their attempts to reach victims and rescue organizations, due to noise/spam unrelated to the disaster and the sheer volume of information shared during a disaster. It is the goal of this study to understand whether machine-based systems trained by humans can identify relevant images shared during the Hurricane Harvey disaster through deep learning methods. We used Twitter's 1% random 'Spritzer' sample retrieved directly from Twitter's streaming API to collect all tweets during our study period and then extracted a sample of Hurricane Harvey-related tweets with the following

* Corresponding author.
E-mail addresses: brett.robertson@utexas.edu, (B.W. Robertson), Dhiraj.Murthy@austin.utexas.edu, (D. Murthy), rothsmith@utexas.edu, (W.R. Smith), keri.stephens@austin.utexas.edu. (K.K. Stephens).

http://dx.doi.org/10.1016/j.pdisas.2019.100030

keywords: 'hurricane', 'harvey', 'hurricaneharvey', 'harveyhouston'. Our sample consists of 17,483 images that were posted within all extracted tweets before (August 17, 2017 to August 25, 2017), during (August 26, 2017 to September 1, 2017) and after Hurricane Harvey (September 2, 2017 to September 17, 2017). 1128 (approximately 6.45%) of the images were randomly selected and were hand-coded by two authors using a theoretically-derived codebook [20] to content analyze the type of need requested and reported. Ultimately, we argue that images produced on public social media during a natural disaster have value to understand how aid and relief can be facilitated, but finding genuine content from disaster victims is not straightforward and involves various complexities. Some of these challenges might be able to be identified and categorized by deep learning methods (particularly around whether an image may signify a case of needing assistance), whereas other attributes might be better served by human interpreters. Our study seeks to explore which challenges are better suited to machines and humans.

2. Social media in disasters

Social media plays some role in most natural and man-made disasters – whether that is around emergency services or soliciting aid – and this use is well established in the literature [29]. However, as social media differ and change in terms of access and features, users have come to use social media for varied purposes during disasters, depending on the type of disaster and social context [25]. Extant literature suggests that citizen use of social media focuses on warnings, response activities, and the quick dissemination of information [48]. Research has demonstrated that the public uses social media as a resource to obtain information during a disaster [51]. This follows larger trends that indicate social media is often an integral, if not prime, source of many people's news mix in developed countries. One reason is that social media often provide access to updated information at a faster rate than traditional news sources. Specifically, Sutton, Palen & Shklovski's [43] early research on the 2007 California wildfires demonstrated that the public used social media because perceptions of official sources and emergency management agencies were that these organizations were not providing important information.

Murthy [24] coins this desire as part of an "update culture," where citizens use social media to stay up-to-date during a disaster. Murthy stresses that these disasters are not all natural ones, but that the proliferation of social media has given rise to how these platforms can be used in manmade disasters as well. Specifically, Potts [32] used the London bombings in 2005 and the terrorist attacks in Mumbai in 2008 as case studies to explore initial social media use in these manmade, terrorist disasters. Further work has looked at social media response after the Sewol ferry disaster in South Korea [52] and campus safety emergencies [40,53]. Together, these studies suggest that social media has been broadly used in a myriad of disaster types, though Murthy [24] stresses that studying the specific platform may shed more light into how these tools can be leveraged.

2.1. Twitter and disasters

Ultimately, Twitter has become a key social medium during disasters. Lachlan et al. [21] state that because Twitter can be used with mobile devices and messages can be shared with large audiences (including those outside one's social network), it can be specifically useful during a natural disaster. In addition to a brief, 280 character limit, users can also link to URLs and images, can retweet others, comment back, and 'like' a tweet. Sutton et al. [44] found that retweets of messages during a 2012 wildfire in Colorado were more likely to take place when the content was advisory, demonstrated a sense of urgency, and had clear sentence structure. Hurricane Sandy in 2012 was one of the first major natural disasters in which scholars analyzed Twitter content. Murthy and Gross's [25] study found that Twitter users often posted images and location-check-ins during Hurricane Sandy, activities which became particularly accentuated as the storm made landfall. Spence et al. [39] also noted that over time, the government-promoted hashtag, "#Sandy," was found by users as unsuitable for discerning useful information as the hashtag was often overshadowed by irrelevant content.

At the time of Hurricane Harvey in 2017, Twitter had 2.46 billon users. Harvey was viewed as a unique natural disaster in which Twitter and other social media provided citizens with a platform to communicate urgent information quickly that ultimately led to life-saving rescues for those who were flooded [41]. In that same study, Stephens et al. [41] discovered that social media posts during Hurricane Harvey served as public service announcements (PSAs), informing both those affected and those outside the storm with important updates. However, determining what social media posts were actually relevant was quite difficult for those seeking information [38]. Similar to Lachlan et al.'s [21] findings, the vast number of tweets related to the actual disaster of Hurricane Harvey varied. Mouzannar et al. [22] stated that much of the information shared during the storm was not usable, and emergency responders who had to manually mine posts for important information often had to sift through *exhausting* amounts of irrelevant content.

2.2. 'Aid' and 'need' on social media

The 2010 earthquake in Haiti brought to light how Twitter can be used for disaster outreach and fundraising on a global level [23], and various humanitarian agencies have used Twitter as a tool for aid mobilization. For example, Frank [7] discusses various organizations using Twitter for cross-promoting ways to make charitable donations after the Haitian earthquake. Further, David, Ong, and Legara [5] argue that besides fundraising, Twitter allows users to build global awareness of a natural disaster. For example, during the 2013 Haiyan typhoon, photos of hard-hit areas were being tweeted by relief workers on the ground, and then shared globally. Thus, the global public was able to be engaged, even if they were not directly affected [5].

In a tsunami evaluation commission report, Telford et al. [45] suggest that the information shared immediately after a disaster is the most valuable for both recovery and future planning. If obtainable, high quality information allows both emergency services and local responders to provide a better emergency response. Yet, according to Goyet and Morinière [8], when information is not provided in a timely manner, a lack of information becomes pervasive amongst stakeholders involved in the natural disaster. Paul [30] argues that the major themes that emerge for tweets immediately following a disaster include: requests, reports and reactions. This encapsulates requests for relief work, including basic human needs such as food, water, shelter and medical assistance [46]. This is followed by requests for search and rescue, infrastructure protection, the recovery of lifeline services, and basic information updates about citizens affected by the natural disaster [36]. Reports often included damage to public and private property, crime, and community mood and behavior. Reactions from the community regarding efforts from emergency response officials, or efforts from the community (e.g., volunteers, food providers) were also common. Paul [30] also notes that tweets during a natural disaster are often not related to the needs of emergency services or providing aid, but could be denoted as spam or marketing, spiritual messages asking for prayers, or personal narratives. Murthy and Longwell [27] also noted that various spam websites were shared at significant levels following the 2010 Pakistan floods.

From a volunteer and nonprofit perspective, Guo and Saxton [9] argue that Twitter serves as a powerful communication tool for social change, especially in educating the public. However, Twitter was viewed as less of a tool for mobilization but more so for providing information to stakeholders and building an online community that could be later called to action. In their analysis of outreach organizations after the 2010 Haitian earthquake, Gurman and Ellenberger [11] found that these same organizations missed opportunities to extend the reach of their message. However, this and other aforementioned research did not specifically understand the role of photos shared during a disaster on Twitter.

2.3. Images on social media

The posting of images and video during natural disasters has become an important part of how these crises are socially experienced and understood [26], though studies of visual content during disasters remains heavily overshadowed by text. The focus on text continues despite the tremendous amounts of information contained in images. Although much research on social media and disasters has focused less on the role of images, these initial studies show that images perform multiple functions and have value. Gupta et al. [10] found that images shared on Twitter during Hurricane Sandy were often spread through retweets. In their research on Instagram images shared during that same hurricane, Murthy et al. [26] argue that images emphasized how people experienced the disaster firsthand, and these images reflected the vantage point of disaster victims rather than official responders. These user-produced images were often shared much faster than what journalists were able to report. Their study was novel given that Hurricane Sandy was the first major natural disaster where Instagram was used. Given that Hurricane Harvey was a unique disaster from a social media viewpoint, we follow Murthy et al.'s advice "to develop ways of tackling these obstacles" for future crises that are socially experienced on Twitter (p. 129).

3. Machine and deep learning methods in natural disasters

Much work has been done regarding classifying Twitter data as signal or noise (i.e., relevant or non-relevant, respectively). For example, the Artificial Intelligence for Disaster Response (AIDR) system performs automatic classification of signal versus noise. Though it is designed for tweets, it is *specifically a text only system*. AIDR uses machine learning methods that are trained on human labeled-data. This has produced quite impressive results. Specifically, AIDR's accuracy has been reported at 80% for identifying relevant tweets during the 2013 Pakistan earthquake [15]. Imran et al. [16] also leveraged machine learning to understand how individuals extracted valuable "information nuggets," which are defined as brief, self-contained information items that could be deemed relevant to man-made disaster response.

Tweedr is another machine-based pipeline that uses tweets to extract actionable information for disaster relief workers [3]; this tool uses classification, clustering and extraction. Recent work studying the use of social media used during Hurricane Harvey by O'Neal et al. [28] employs supervised learning methods, specifically with images, as AIDR and Tweedr are text-based. O'Neal et al.'s [28] study, however, is not focused on Twitter, but seeks to evaluate the use of supervised learning based on samples of private social media data. Deep learning methods are not evaluated, nor are training sets developed from noisy, public social media platforms. However, their work suggests machines are likely able to learn from human knowledge and leverage this to classify the basic features of images by categories (e.g. rescuee and rescuer). In Elbanna et al.'s [6] work, a series of workshops with first responders revealed the need for machine learning to be applied to social media data in meaningful ways, given that disaster victims are increasingly using social media as their first lifeline in crises. In turn, governmental agencies who manage disasters are interested in novel methods to capture emerging social media use during disasters — and machine learning may be one way to classify those who need help.

4. Research questions

Based on our review of the literature, the need suggested by Elbanna et al. [6], and the call to action proposed by O'Neal et al. [28], we propose the following research question:

RQ1: Can we achieve high, real-time accuracy and classification rates of images – ignoring included text - posted to Twitter during a natural disaster for aid-and-need using deep learning machine methods with a human in the loop? If so, can this be done on harder classification tasks such as the time period an image corresponds to or the state of urgency an image represents?

Our next goal is to gather insights from manual coding of images posted to Twitter during a natural disaster and compare those to developing training sets for deep learning methods. Therefore, we propose the following research question:

RQ2: What are the unique values of human coding inputs when deployed alongside deep learning methods?

5. Method

5.1. Data collection

To collect data, we used Twitter's 1% random 'Spritzer' sample retrieved directly from Twitter's streaming API. We studied tweets from August 17, 2017 to September 17, 2017 and extracted from our Spritzer sample all Hurricane Harvey-related tweets with the keywords: 'hurricane', 'harvey', 'hurricaneharvey', 'harveyhouston'. From these tweets, we extracted all the non-video media-related links to retrieve images. Duplicate images were removed by computing an MD5 checksum for each image, which is an "algorithm that is used to verify data integrity through the creation of a 128 bit message digest from data input (which may be a message of any length); the product is claimed to be as unique to that specific data as a fingerprint is to the specific individual" ([1], p. 132). An MD5 checksum "scheme guarantees storing exactly the same file only once and easily identifying duplicates or near duplicates (accounting for images in various formats, at different resolutions and with minor modifications such as some watermarks)" ([50], p. 3).

The resulting total number of images was 23,692. 17,483 images remained after duplicates and empty images were removed. To develop the training dataset for our deep learning classification study, we randomly sampled 1128 images (approximately 6.45%) and human coded these images using a rubric with 10 questions [see human coding information in 4.2 and Appendix A for the rubric].

5.2. Deep learning pipeline

In this section, we describe the methodology used to classify images by time period and urgency. A high-level overview of our methodology is illustrated in Fig. 1.

5.2.1. Transfer learning for feature extraction

Transfer learning refers to application of knowledge gained by solving a prior problem to a new, but related problem. The effectiveness of classic deep learning methods, like Convolutional Neural Networks, is limited by the size of the training set [13]; transfer learning offers the benefits of these deep learning methods while not requiring a large training set. The high dimensionality of images, typically represented as a matrix of pixels where each pixel has a value for its red, green, and blue elements, can be challenging for traditional machine learning methods to interpret. Instead of feeding raw images into models, images can be fed into convolutional neural networks to reduce the dimensionality of images. These networks use convolutional and pooling layers to extract features, or "feature vectors" before making a classification or regressive prediction. These feature vectors can be understood as the collection of nonlinear features, such as edges, shadows, and areas of interest that fully describe the original image. In this way, pre-trained convolutional neural networks trained on large, diverse datasets can be used as feature extractors for other machine learning tasks [12].

For the purpose of this study, we used VGG-16, a popular convolutional neural network traditionally used to classify images into categories of objects, as a method of feature extraction [37]. Instead of classifying candidate images into categories, we collected the output from the second-to-last layer of VGG-16 before classification and treated this output as a feature vector representing each image.

5.2.2. Multi-layer perceptron models (MLPs)

After extracting the feature vectors for each model, two multilayer perceptron networks were constructed to classify images by time period

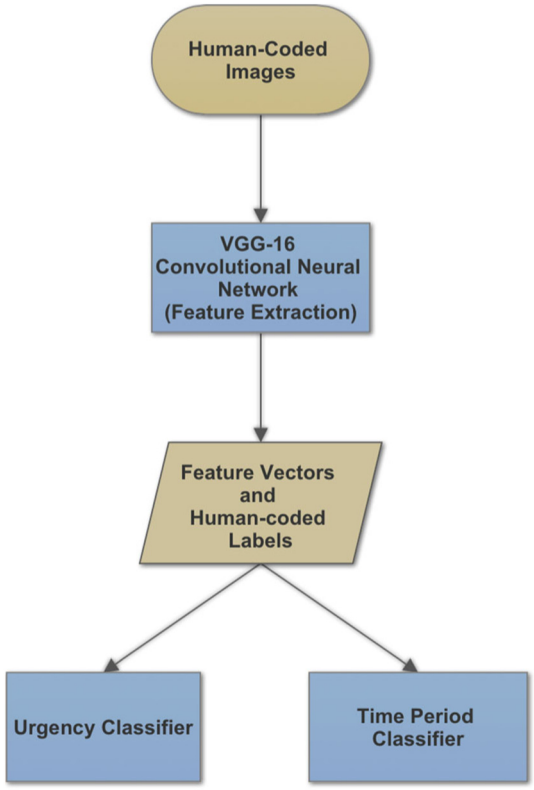

Fig. 1. High-level overview of image classification pipeline.

and urgency for the labeled images. For each model, the image and label pairs were randomly split into a training and validation set, where each set represented a stratified random sample of images and their corresponding labels. Ultimately, the training set constituted 67% of the total images and the validation set consisted of the remaining 33% images.

Each model was evaluated against the categorical cross-entropy loss function [54] and its classification accuracy. Error minimization was achieved using the Adam optimizer, which has been shown to converge faster compared to traditional methods like stochastic gradient descent [18]. To combat the effect of class imbalance on model training, we scaled the penalty on misclassification according to the proportion of samples which contained that label. Furthermore, each model was trained with early-stopping, which halts model training once an increase in training accuracy is accompanied by a significant decrease in the validation accuracy of the model.

Each feed-forward network consisted of three dense layers, with each subsequent layer having fewer nodes than the previous layer. Each layer contained a sigmoidal activation function to normalize each layer's output to values between 0 and 1. Furthermore, each dense layer was succeeded by a dropout layer, which randomly reset the weights of a subset of nodes in that layer. Finally, the output of the last dense layer is passed through a softmax activation layer so that the final output of each network represented a classification probability for each of the relevant classification categories. Fig. 2 describes the architecture of the feedforward networks used to classify each attribute.

6. Results of transfer learning

6.1. Time period classifier

The first classifier constructed predicted the time period defined as pre-storm (August 17, 2017 to August 25, 2017; $n = 124$ images), landfall (August 26, 2017 to September 1, 2017; $n = 735$ images), and Harvey's aftermath/immediate cleanup (September 2, 2017 to September 17, 2017; $n =$

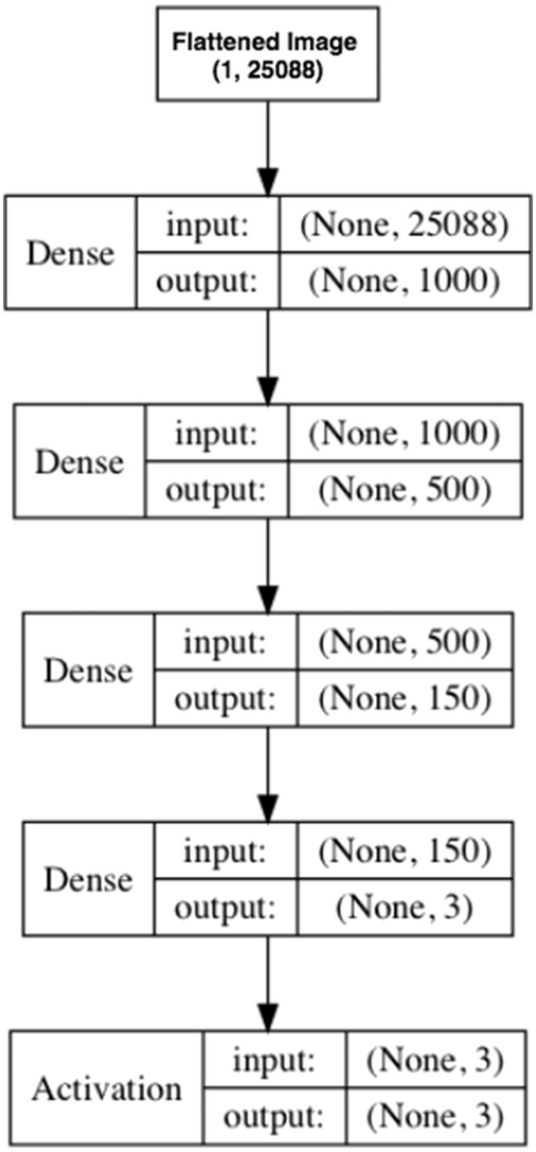

Fig. 2. MLP classifier architecture.

269 images). The time period classifier ultimately reached a training accuracy of 0.8954, a training loss of 0.3325, a validation accuracy of 0.6461, and a validation loss of 1.0172 in 5 epochs, defined as a full training cycle on the training set, in batches of 32. Fig. 3 illustrates the changes in training and validation accuracy and categorical cross entropy for each epoch of training. Prima facie, the high performance of the time period classifier seems to indicate that deep learning methods based on a small sample of images could be quickly deployed to identify whether the post date of an image on social media likely corresponded with not only when the image was taken, but, whether it content-wise corresponded to before, during, or after a disaster hit.

Fig. 4 shows a heat-mapped confusion matrix for the time period classifier on the validation data.

However, as the confusion matrix illustrated in Fig. 4 indicates, this is likely due to the classifier guessing time period 1 ‘landfall’, rather than actually developing human-like neural pathways for recognizing what stage the disaster occurred at. Specifically, despite the time period classifier's relatively high accuracy, the confusion matrix for the time period classifier suggests the model's tendency to predict time period ‘1’ regardless of an image's feature vectors, despite our attempt to penalize such misclassifications by scaling the penalty in accordance with the relative frequency of the time period labels. It is likely that the classifier simply predicts time

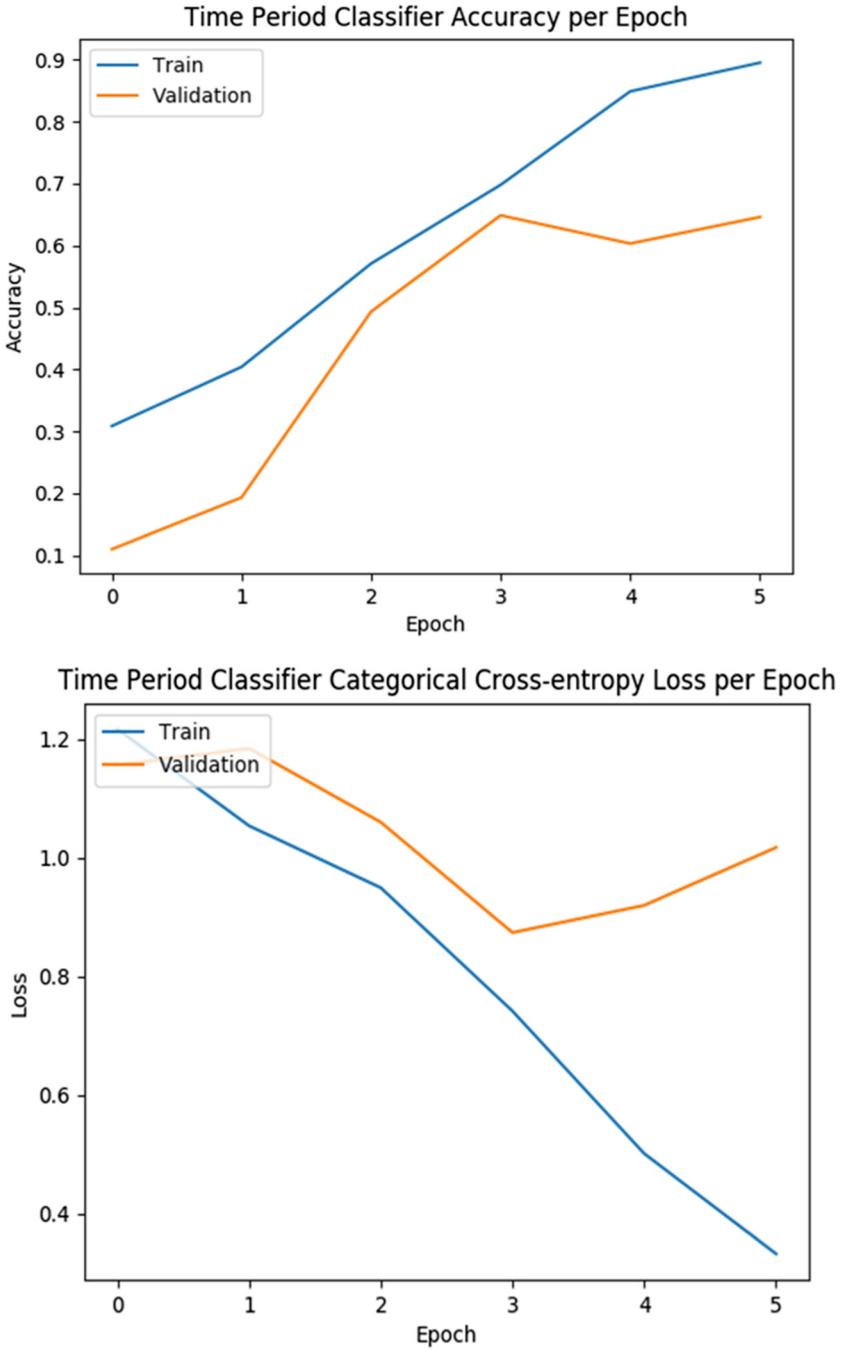

Fig. 3. Time period classifier accuracy and loss by epoch.

period '1' because the optimal accuracy for the classifier is achieved when the majority of images are classified as the most frequent time period. Of the 89 images in the validation set from time period '2', only 6.7% were classified correctly; likewise, of the 41 images from time period '0', only 7.3% were classified correctly. The high performance of our model could also be explained by class imbalance rather than the extraction of useful feature vectors. The implication of this for the study of images posted during disasters is that people are still sharing and communicating images more frequently during, rather than before or after a disaster.

The frequency of images posted during the disaster is significantly larger than images posted before or after. Fig. 5 provides a frequency distributed from the images coded. This may suggest that, despite the restrictions on power and cellular service imposed during a disaster, images remain a valuable and important medium for people to communicate during natural disasters [38]. Furthermore, the time period classifier's poor performance when classifying time period may suggest that, in order for deep learning classifiers to be successful, a large dataset of labeled image is needed. In scenarios where classification of these images is time-sensitive, this suggests that deep learning models for time period classification may require more human labeling than traditional methods of image classification.

6.2. Urgency classifier

The urgency classifier was trained to predict the urgency of an image into different levels. Saldana [35] suggests that it is important to identify the level of importance of a social media post by adding a magnitude coding to the coding scheme. Here, images were ranked in terms of importance to

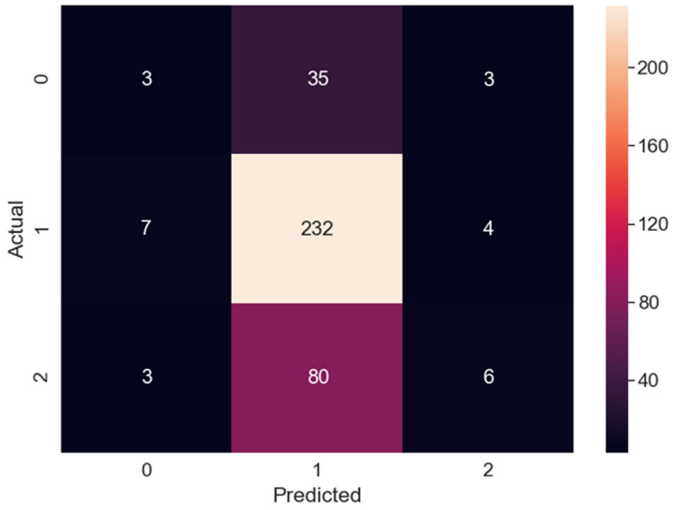

Fig. 4. Confusion matrix for time period classifier.

Hurricane Harvey (4 = highly urgent, 3 = moderately urgent, 2 = somewhat urgent, 1 = not urgent, 0 = spam/unrelated to Hurricane Harvey), similar to Iakovou and Douligeris [14] recommendations on severity of a hurricane. The urgency classifier ultimately reached a training accuracy of 0.6768, a training loss of 0.8868, a maximum validation accuracy of 0.3038, and a validation loss of 1.7252 in 5 epochs, where each epoch represents a single pass of all training data through the network in batches of 32. Fig. 6 illustrates the changes in training and validation accuracy and categorical cross entropy for each epoch of training.

Fig. 7 shows a heat-mapped confusion matrix for the time period classifier on the validation data.

Similar to the time period classifier, the confusion matrix for the urgency classifier suggests that the model tends to favor predicting that an image belongs to urgency levels of '0' and '2'; 71.4% of the model's predictions belong to one of these categories, whereas 64.4% of the validation set consists of images labeled as either '0' or '2'. Table 1 summarizes the empirical results for both of the trained classifiers.

Both classifiers indicate issues as revealed by Confusion Matrix testing (see Figs. 4, 7). This is not uncommon with image classification testing; however, this suggests some instabilities with our models, particularly as it relates to class imbalance. In the case of the time-period classifier, it is possible that additional data, from the less-frequently occurring classes could address the issue posed by class imbalance. While the urgency classifier displays many of the same issues as the time period classifier, it is possible that filtering out the 'spam' images, marked as '0', prior to model training, could improve the performance of the classifier. Despite these results, we believe that coding a larger random sample of images, coupled with filtering out the images labeled 'spam' prior to model training could improve the performance of the classifiers.

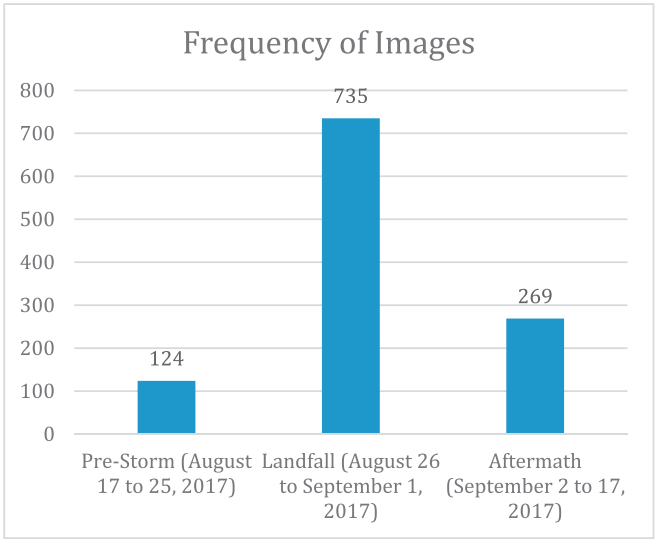

Fig. 5. Frequency of images by time period.

7. Human coding results

To help answer RQ1 and RQ2, human coding content analysis was used, as both a way to create a training model, and for its own unique qualitative utility. In this section, we detail our qualitative method and results of this study. For this analysis, 1128 images (from the 17,483 images) were randomly selected to be manually coded by two coders (approximately 6.45%).

A preferred method is to do a content analysis by creating a coding schema and manually code to evaluate tweets—the unit of analysis [4,20]. In our study, the unit of analysis is images posted on Twitter. Understandably, images are far more challenging than text and due to the large volume of images needed to interpret, our coding methods focused on broader 'motif' categories, a common practice in disaster-related social media research [26,49]. Each image was coded using a closed-deductive codebook, and the coding framework was theoretically derived from previous literature [20] that represented the following variables: the urgency of the image [14,35], type of image posted [30], a description of the image [30], and the type of motif the image represented [26]. The categories we used for coding were designed to relate to all three phases of Harvey: pre-storm, landfall, and Harvey's aftermath and immediate cleanup. See Appendix A for codebook.

7.1. Coder training and intercoder reliability

Prior to coding, the authors used the literature to draft an initial coding framework. Two coders coded the same dataset, and average Cohen's kappa was 0.937 across coded categories, signifying very strong coder agreement. The coders held a meeting prior to coding to discuss the codebook and trained with a sample dataset for practice together. Throughout coding, the coders met twice to discuss operationalization and trends that emerged from the coding process.

7.2. Frequency data

Drawing from Stephens & Malone's [42] content analyses methodology, we present the results of the content analyses as frequencies to provide a grounding for the deep learning methods used in study one of this manuscript. It appears almost one third (exactly 31.6%) of images shared on Twitter were not related to the Hurricane Harvey disaster (e.g., spam or noise), while 31.2% of images shared were seen as 'somewhat urgent' in the perception of an image's urgency. See Fig. 8. We can perhaps interpret these results by demonstrating that highly urgent information was not necessarily disseminated through Twitter images during Harvey (7.7% of Twitter images were coded as highly urgent). Research (see [38,41]) explains that highly urgent information (including addresses and phone numbers for rescues) during the disaster was often shared through private social media feeds, not public.

Using Paul's [30] typology of social media posts shared during a disaster, we find that many posts shared during Hurricane Harvey were reports (n = 740), including reports of damage, reporting community behavior, and reporting news coverage (Fig. 9). Note that the sample size does not add up to 1128, as some images did not fit the typology, and other images could fit more than one code. We also see requests (n = 26) were not shared as much as reports and reactions. Requests included immediate help and rescue, material support, medical assistance, or simply for information. Like the 'highly urgent' finding above, it makes sense that perhaps Twitter was not used during Hurricane Harvey for requesting aid and need in

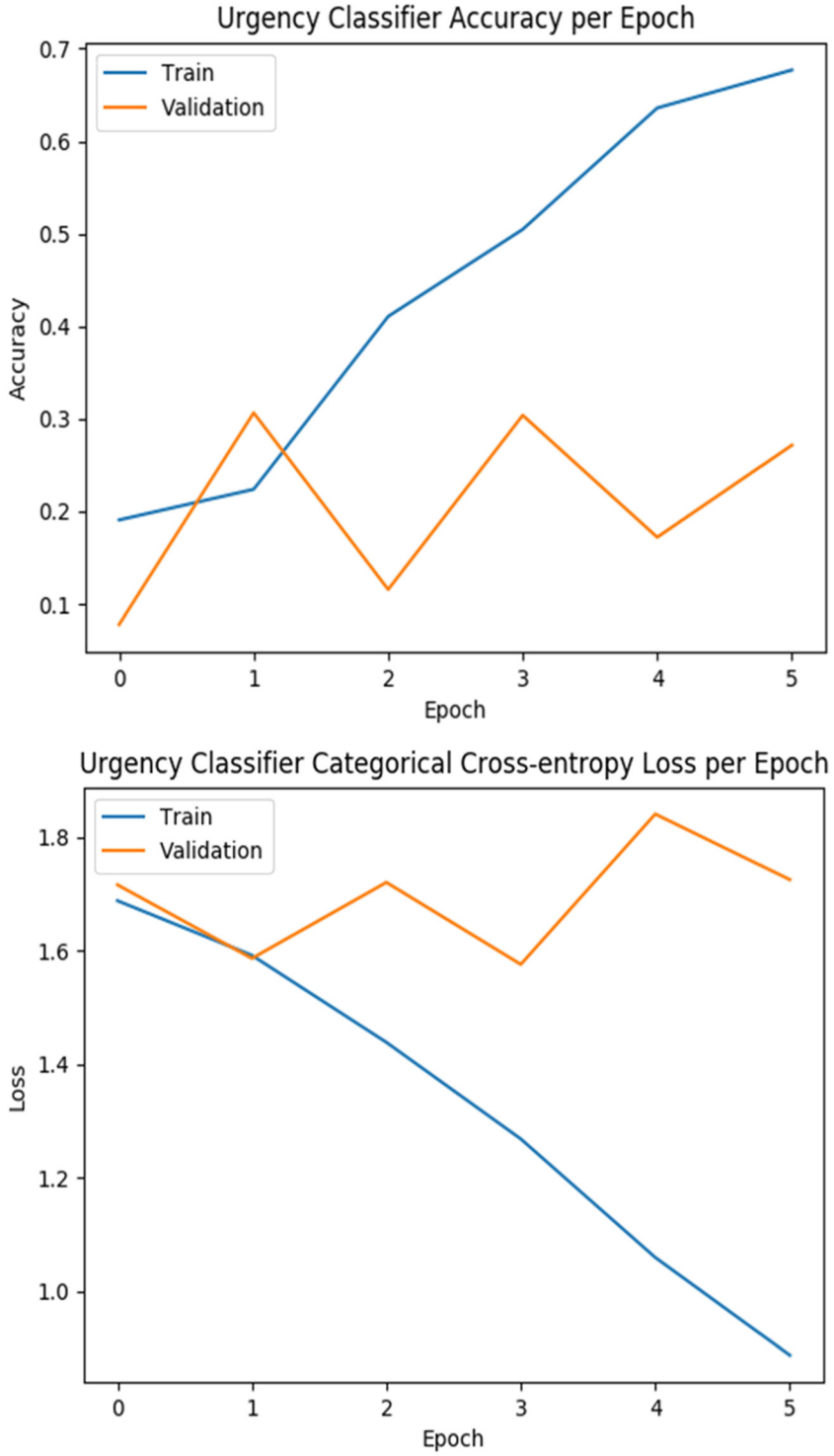

Fig. 6. Urgency classifier accuracy and loss by epoch.

situations requiring highly urgent and time sensitive assistance, following in line with Smith et al.'s [38] finding that private social media platforms (e.g., neighborhood Facebook groups and Nextdoor community pages) were used more frequently for the purposes of seeking and providing rescue to teams with high-water rescue boats.

Finally, drawing from the work by Murthy et al. [26] looking at images crossed-posted to Twitter and Instagram during Hurricane Sandy, we employ their same coding framework to image motifs shared during Hurricane Harvey. Murthy et al. [26] argue that using motif categories allows us to understand the basic social experience of disasters.

Table 2 illustrates the frequency of codes applied to images. The majority of the motifs can be taken at face value, however, the "MACRO" motif was used to denote 'image macros', images that have a "picture superimposed with text with a specific purpose of being funny" ([26], p. 119). The motif of "GEAR" included any equipment or supplies (e.g., boats, trucks, etc.). Aggregated frequencies are important because they help to tell alternative stories of Harvey, rather than simply mainstream accounts of the disaster. Specifically, there is a high frequency of the motifs OUTSIDE (defined as images depicting the built environment, nature, or spaces/places not indoors), PEOPLE (defined as images depicting

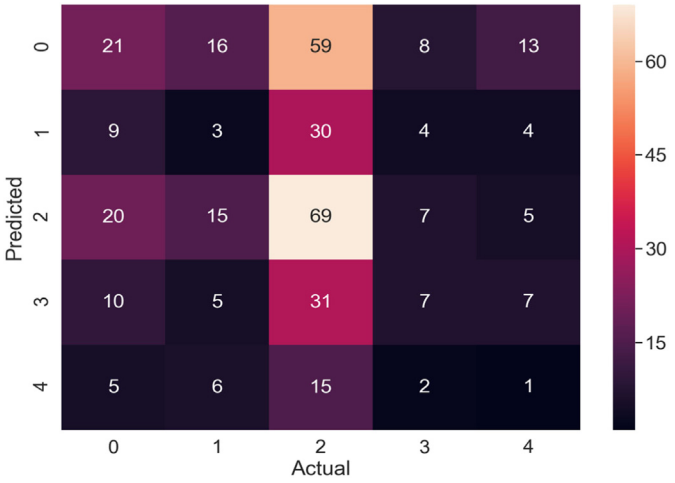

Fig. 7. Confusion matrix for the urgency classifier.

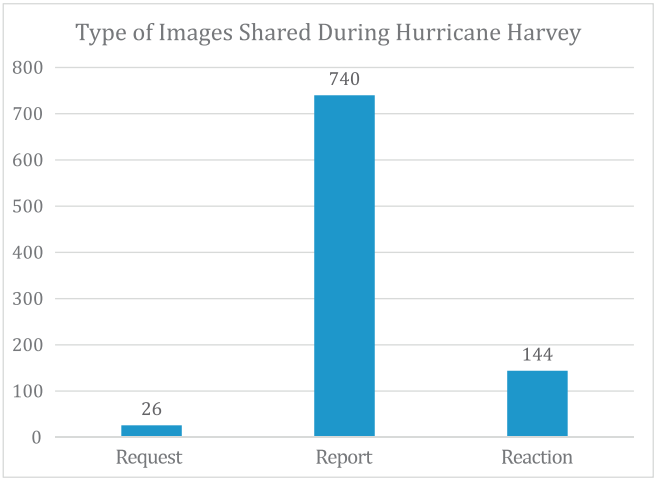

Fig. 9. Type of image (n.b. sample size does not total 1128 as some images did not fit the typology and some images fit more than one typology).

Table 1
Empirical results of classifiers.

Classifier	Training loss	Training accuracy	Validation loss	Maximum validation accuracy
Time period	0.3325	0.8954	1.017	0.6461
Urgency	0.8868	0.6768	1.7252	0.3038

people, not inclusive of cartoon depictions of people; inclusive of selfies, individuals, and groups of people) and DAMAGE (defined as images depicting storm-related damage to the built environment). We interpret this as individuals' desire to document the disaster experience through showcasing weather-related conditions, instances of flooding, and high-water rescues. This is in juxtaposition to the low frequency of RELIEF (defined as images depicting relief efforts and relief campaigns, inclusive of screenshots of relief campaigns). We also have 'OTHER' with a relatively high frequency. As the two coders met to discuss findings from the dataset, they agreed that a large number of the images includes maps, weather radars, or wind and rainfall predictions ($n = 168$ that included "map"). "MAPS," as an emerging motif category in our data, give credence to the idea that during this disaster, people were more concerned about the prediction of weather and changes in the forecast. This is an important change in the image corpora of Hurricane Sandy versus Harvey. Together, these image frequencies help sketch an outline of the types of narratives that unfolded during Hurricane Harvey. Taken together, the motifs paint a picture of the disaster from the lived experiences of the individuals who experienced it.

8. Discussion

In this paper, we present a framework for feature extraction using the VGG-16 convolutional neural network and construct multilayer perceptron classifiers for classifying the urgency and time period for a given image. This framework was created through a qualitative deductive coding schema, in which the qualitative results (presented as frequencies) describe the unique disaster experience through the eyes of photos. Of course, the

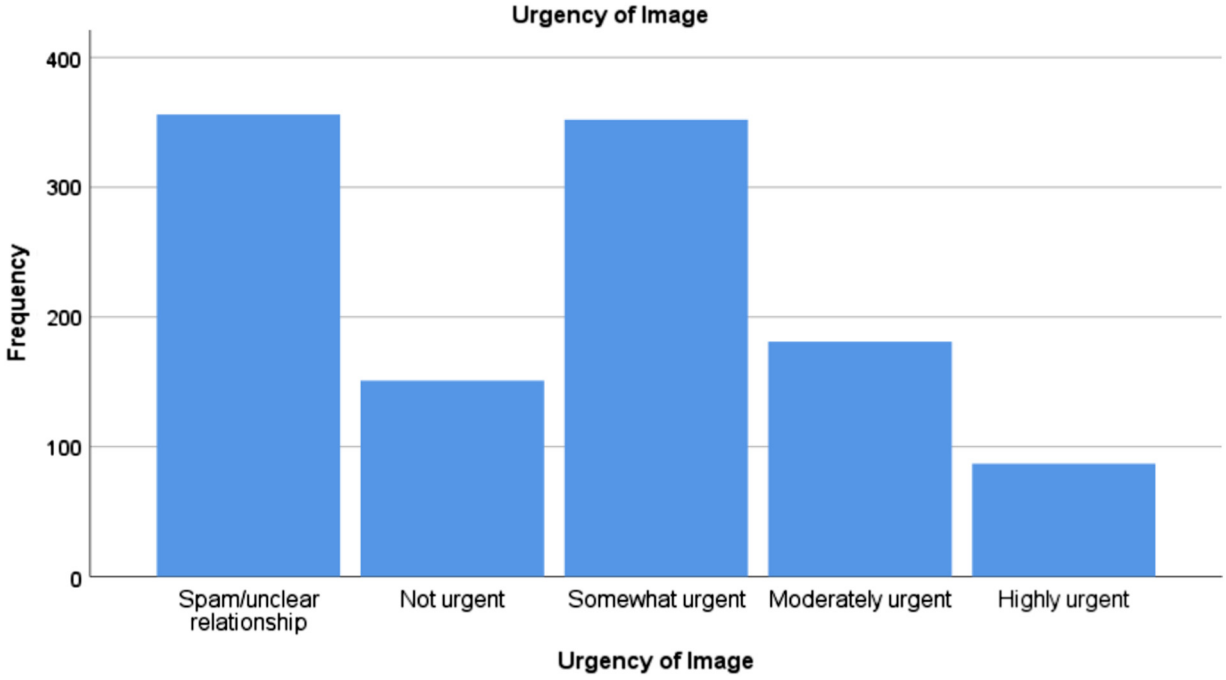

Fig. 8. Urgency of image to Hurricane Harvey.

Table 2
Frequency of coded image motifs.

	Frequency	Percent
Ad	5	0.25
Animals	56	2.85
Damage	295	15.03
Drink	15	0.77
Food	20	1.02
Gear	99	5.05
Macro	66	3.40
Other	404	20.59
Outside	492	25.07
People	428	21.81
Relief	82	4.18
Total	1962	100% (not exact due to rounding)

images produced during a natural disaster on social media have value to understand how aid and relief can be facilitated, but finding genuine content (and not noise), is not straightforward and we detail our approach to address these challenges and provide initial results for our deep learning image classification methodology.

The results obtained by using transfer learning to extract features to feed to a simple multi-layer perceptron model indicate both models learned the relationship between feature vectors from the VGG-16 model and each image's urgency and time period beyond random guessing; however, the tendency of both models to grossly misclassify images towards more frequent labels suggests a lack of robustness for our trained models. We suggest two possibilities for the weak performance of our classifiers: class imbalance and dataset size.

8.1. Class imbalance

The descriptive analysis on the frequency of labels for time period and urgency coupled with each model's tendency to favor labels which occur more frequently suggest class imbalance as a potential explanation for the weak performance of our models. While solutions to the class-imbalance problem exist on both data and algorithmic levels [19], the class imbalance in the dataset used for our study could be explained by the inherent bias in human-coding. This bias may be present when images are more likely to be classified as 'spam/not relevant' and 'somewhat urgent' due to the inclusive, catch-all wording of these categories.

8.2. Dataset size

Recent literature suggests that transfer learning for feature extraction can be useful when the number of available training samples is low [55]. However, the noisy, diverse set of images characteristic of those posted on Twitter can be difficult to capture in such a small number of samples, even with transfer learning. It is possible that the sample of images collected in this study do not adequately encapsulate the patterns and relationships between the feature vectors collected from VGG-16 and therefore the deep learning models constructed are unable to extract meaning from the feature vectors and their corresponding labels.

In terms of qualitative results, the use of qualitative image coding was important, not only for creating the training model, but on its own merit, as scholars argue that image content on social media goes beyond how people experienced the disaster firsthand, and these images reflected the vantage point of disaster victims and official relief and rescue organizations. Our results demonstrate that highly urgent and time sensitive images (including requests for immediate help, material support, and information) were the least shared on Twitter. This makes sense given the context of Hurricane Harvey, where those seeking help and those providing help often used private social networks for rescue efforts. Therefore, the utility of tweets for urgent aid requests during disasters might be overemphasized in existing literature. We also expand upon the motif framework provided by Murthy et al. [26] and find that maps emerged as a motif category in our dataset. This indicates that there are some changes in the types of images being posted during Hurricane Harvey versus Hurricane Sandy. Computational work with future hurricane data would be useful to chart changes in image motifs over time and whether particular types of disasters (earthquake versus hurricane) caused major changes in motif types and their frequency.

The tendency of both classifiers to act as 'lazy' classifiers and predict only the most frequent categories may suggest that, in their current state, these deep learning classifiers require additional tuning, data, and preprocessing in order to truly be effective. For example, it is plausible that filtering out the spam and irrelevant images from the dataset the urgency classifier could drastically improve the classifier's ability to learn the complex relationships between the feature vectors and their corresponding urgency and time period labels. In Table 3, we present a summary of the comparison between our machine learning and human-coded results.

9. Conclusion

Overall, it is crucial to continue probing how machine learning can aid in disaster response, particularly considering that the 9-1-1 system in the United States can become overloaded with calls for help. The scale of social media data relevant to studying disasters is only likely to grow. This study is novel by taking a first step at investigating whether or not deep learning machine methods can filter through the noise on social media and identify authentic calls for help or urgent situations during a disaster.

This study is one of the first studies to combine qualitative methods and results using a traditional content analysis with 'deep' machine learning methods. While both methods inform each other in our study, they also provide utility on their own. After reviewing the literature, our team found that no study has used these two methods together. Although our classifiers did not perform as expected, our research opens the door for new interdisciplinary methods to be used in future disaster research.

Our study does have some limitations. Specifically, we use 1128 randomly selected images (approximately 6.45%) from a larger corpus of 17,483 images. While smaller data sets have been found to be acceptable for transfer learning methods, the classifiers still performed weaker than anticipated. Coding a larger percentage of images may be more useful for both transfer learning and traditional content analyses, and our future work seeks to address this. This analysis also did not include the use of

Table 3
Comparison of machine learning and human coded results.

Machine learning results	Human-coded results
• Successfully used transfer learning to extract features to feed to a simple multi-layer perceptron model indicate both models learned the relationship between feature vectors from the VGG-16 model and each image's urgency and time period beyond random guessing. • However, the tendency of both models to grossly misclassify images towards more frequent labels suggests a lack of robustness for our trained models. • If images are time-sensitive, a deep learning model for time period classification may require more human labeling than traditional methods of image classification. • Coding a larger random sample of images, coupled with filtering out the images labeled 'spam' prior to model training could improve the performance of the classifiers.	• Close to one third (exactly 31.6%) of images shared on Twitter were not related to the Hurricane Harvey disaster, while 31.2% of images shared were seen as 'somewhat urgent' in the perception of an image's urgency. • Many posts shared during Hurricane Harvey were reports ($n = 740$), including reports of damage, reporting community behavior, and reporting news coverage. • Many images shared represented thematic motifs of outside, people and damage. Maps emerged as a motif to depict weather-related conditions. • These human codes serve, not only on their own qualitative/content analysis merit, but work hand-in-hand within the image classification of the deep-learning pipeline.

rumors or misinformation, but this was common during Hurricane Harvey (see [41] for examples of how citizens dealt with misinformation). Finally, our work only addressed images shared on Twitter. Other social media platforms, including private Facebook groups, Nextdoor, Instagram, WeChat, Mastodon and Weibo are also likely to be fruitful venues of data for future work seeking to understand image-based sharing.

Our results have particular significance to hurricane events and we are unsure how generalizable our findings are to other disasters – particularly outside the contiguous U.S. As Palen and Hughes [29] lament, lessons learned from one kind of emergency may not be applicable to others, even when the medium stays the same. Although Hurricane Harvey and its effect on the greater Houston area were unique, our work provides a new framework that could be readily drawn upon when disaster strikes.

Declaration of Competing Interest

None.

Acknowledgments

Funding: This work was supported by a grant from the National Science Foundation [award # 1760453] RAPID/The Changing Nature of "Calls" for Help with Hurricane Harvey: 9-1-1 and Social Media. Any opinions, findings, and conclusions or recomendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Theoretical codebook

Date/Image

- Date of Image
- Image Number

Time Period

- 1 = immediately prior: August 17 to 25
- 2 = during: August 26 to September 1, 2017
- 3 = after/immediate clean-up efforts: September 2 to September 17, 2017

Relevancy to Hurricane Harvey

- 0 = no
- 1 = yes
- 2 = uncertain

Urgency of Image [14,35]

- 4 = highly urgent
- 3 = moderately urgent
- 2 = somewhat urgent
- 1 = not urgent
- 0 = spam/unclear relationship to disaster

Type of Image in a Disaster [30]

- 1 = Request
- 2 = Report
- 3 = Reaction

Description of Type of Image [30]

- 1 = request for material support
- 2 = request for medical assistance
- 3 = request for information
- 4 = request for immediate help/rescue
- 5 = report of damage
- 6 = reporting community behavior
- 7 = report of news coverage
- 8 = reaction from community
- 9 = reaction from official sources

Motif [26]

- 1 = ad
- 2 = animals
- 3 = damage
- 4 = drink
- 5 = food
- 6 = gear
- 7 = macro
- 8 = outside
- 9 = people
- 10 = relief
- 11 = other

Image attributes/keywords [open ended]

- map, meme, cartoon, celebrity, water, house, street, bridge

Text attributes/keywords [open ended]

- percentage, address, request, phone number

References

[1] Ahmad MA, Alshaikhli DI, Alhussainan SO. Achieving security for images by LSB and MD5. J Adv Comput Sci Technol Res 2012:127–39.
[2] American Red Cross. New study shows more reliance on mobile apps in disasters. Retrieved from http://www.redcross.org/news/article/New-Study-Shows-More-Reliance-on-Mobile-Apps-in-Disasters; 2012.
[3] Ashktorab Z, Brown C, Nandi M, Culotta A. Tweedr: mining twitter to inform disaster response. In: Hiltz SR, Pfaff MS, Plotnick L, Robinson AC, editors. Proceedings of the 11th International ISCRAM Conference; 2014 [University Park, Pennsylvania, USA].
[4] Bruns A, Burgess JE, Crawford K, Shaw F. #qldfloods and @QPSMedia: crisis communication on Twitter in the 2011 south east Queensland floods. Retrieved from https://eprints.qut.edu.au/48241/1/floodsreport.pdf; 2012.
[5] David CC, Ong JC, Legara EFT. Tweeting supertyphoon Haiyan: evolving functions of Twitter during and after a disaster event. PloS one 2016;11(3):e0150190.
[6] Elbanna A, Bunker D, Levine L, Sleigh A. Emergency management in the changing world of social media: framing the research agenda with the stakeholders through engaged scholarship. Int J Inf Manag 2019;47(1):112–20.
[7] Frank T. Social media play part in Haiti's recovery efforts. USA Today. Retrieved from http://usatoday30.usatoday.com/tech/news/2010-02-01-haiti-monitor-social-media_N.htm; 2010, February 1.
[8] Goyet D, Morinière LC. The role of needs assessment in the tsunami response. Retrieved from https://docs.unocha.org/sites/dms/Documents/TEC_Needs_Report.pdf; 2006.
[9] Guo C, Saxton GD. Tweeting social change: how social media are changing nonprofit advocacy. Nonprofit Volunt Sect Q 2014;43(1):57–79.
[10] Gupta A, Lamba H, Kumaraguru P, Joshi A. Faking sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. Proceedings of the 22nd international conference on World Wide Web. ACM; 2013, May. p. 729–36.
[11] Gurman TA, Ellenberger N. Reaching the global community during disasters: findings from a content analysis of the organizational use of Twitter after the 2010 Haiti earthquake. J Health Commun 2015;20(6):687–96.
[12] Hertel L, Barth E, Käster T, Martinetz T. Deep convolutional neural networks as generic feature extractors. Neural Networks (IJCNN), 2015 International Joint Conference on (pp. 1–4). IEEE; 2015, July.
[13] Hosseini H, Xiao B, Jaiswal M, Poovendran R. On the limitation of convolutional neural networks in recognizing negative images. Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on (pp. 352–358). IEEE; 2017, December.
[14] Iakovou E, Douligeris C. An information management system for the emergency management of hurricane disasters. Int J Risk Assess Manage 2001;2(3):243–62.
[15] Imran M, Castillo C, Lucas J, Meier P, Vieweg S. AIDR: artificial intelligence for disaster response. Proceedings of the 23rd International Conference on World Wide Web (pp. 159–162). ACM; 2014, April.
[16] Imran M, Elbassuoni S, Castillo C, Diaz F, Meier P. Extracting information nuggets from disaster-related messages in social media. In: Comes T, Fiedrich F, Fortier S, Geldermann J, Yang L, editors. Proceedings of the 10th International ISCRAM Conference. Germany: Baden-Baden; 2013, May.
[17] King LJ. Social media use during natural disasters: an analysis of social media usage during Hurricane Harvey. Proceedings of the International Crisis and Risk Communication Conference. FL: Orlando; 2018.
[18] Kingma, D. P., & Ba, J. (2014). Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[19] Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Art Intell 2016;5(4):221–32.
[20] Krippendorff K. Content analysis: an introduction to its methodology. Beverly Hills, CA: Sage; 1980.

[21] Lachlan KA, Spence PR, Lin X. Expressions of risk awareness and concern through Twitter: on the utility of using the medium as an indication of audience needs. Comput Hum Behav 2014;35:554–9.
[22] Mouzannar H, Rizk Y, Awad M. Damage identification in social media posts using multimodal deep learning. In: Boersma K, Tomaszewski B, editors. Proceedings of the 15th International ISCRAM Conference, Rochester, NY, May 2018; 2018.
[23] Muralidharan S, Rasmussen L, Patterson D, Shin JH. Hope for Haiti: an analysis of Facebook and Twitter usage during the earthquake relief efforts. Public Relat Rev 2011;37(2):175–7.
[24] Murthy D. Twitter: social communication in the Twitter age. Cambridge: Polity; 2018.
[25] Murthy D, Gross AJ. Social media processes in disasters: implications of emergent technology use. Soc Sci Res 2017;63:356–70.
[26] Murthy D, Gross AJ, McGarry M. Visual social media and big data: Interpreting Instagram images posted on Twitter. Digit Cult Soc 2016;2(2):113–34.
[27] Murthy D, Longwell SA. Twitter and disasters: the uses of Twitter during the 2010 Pakistan floods. Inf Commun Soc 2013;16(6):837–55.
[28] O'Neal A, Rodgers B, Segler J, Murthy D, Lakuduva N, Johnson M, Stephens KK. Training an emergency-response image classifier on signal data. Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA); 2018. p. 751–6.
[29] Palen L, Hughes AL. Social media in disaster communication. In: Rodríguez H, Donner W, Trainor JE, editors. Handbook of disaster research. 2nd ed. Cham, Switzerland: Springer; 2010. p. 497–520.
[30] Paul A. Identifying relevant information for emergency services from twitter in response to natural disaster. Retrieved from https://eprints.qut.edu.au/89220/1/Avijit_Paul_Thesis.pdf; 2015.
[31] Pew Research Center. Social media use in 2018. Retrieved from http://www.pewinternet.org/2018/03/01/social-media-use-in-2018/; 2018, March 1.
[32] Potts L. Social media in disaster response: how experience architects can build for participation. New York, NY: Routledge; 2013.
[33] Rainear, A. M., Lachlan, K. A., Oeldorf-Hirsch, A., & DeVoss, C. L. (2018). Examining Twitter content of state emergency management during Hurricane Joaquin. Communication Research Reports, Advanced online publication, 1–10.
[34] Rhodan M. Please send help: Hurricane Harvey victims turn to Twitter and Facebook. Time 2017, August 30 Retrieved from http://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/.
[35] Saldana JM. The coding manual for qualitative researchers. . 2nd ed.London, UK: SAGE Publications; 2012.
[36] Si X-S, Wang W, Hu C-H, Zhou D-H. Remaining useful life estimation–a review on the statistical data driven approaches. Eur J Oper Res 2011;213(1):1–14.
[37] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv, vol. 1409; 2014; 1556.
[38] Smith WR, Stephens KK, Robertson BW, Li J, Murthy D. Social media in citizen-led disaster response: Rescuer roles, coordination challenges, and untapped potential. In: Boersma K, Tomaszewski B, editors. Proceedings of the 15th International ISCRAM Conference, Rochester, NY; 2018, May.
[39] Spence PR, Lachlan KA, Lin X, del Greco M. Variability in Twitter content across the stages of a natural disaster: implications for crisis communication. Commun Q 2015;63(2):171–86.
[40] Stephens KK, Barrett AK, Mahometa MJ. Organizational communication in emergencies: using multiple channels and sources to combat noise and capture attention. Hum Commun Res 2013;39(2):230–51.
[41] Stephens KK, Li J, Robertson BW, Smith WR, Murthy D. Citizens communicating health information: urging others in their community to seek help during a flood. In: Boersma K, Tomaszewski B, editors. Proceedings of the 15th International ISCRAM Conference; 2018, May [Rochester, NY].
[42] Stephens KK, Malone PC. New media for crisis communication: opportunities for technical translation, dialogue, and stakeholder responses. In: Coombs TC, Holladay SJ, editors. Handbook of crisis communication. Malden, MA: John Wiley & Sons; 2010. p. 381–95.
[43] Sutton JN, Palen L, Shklovski I. Backchannels on the front lines: emergency uses of social media in the 2007 Southern California Wildfires. University of Colorado; 2008; 624–32.
[44] Sutton J, Spiro ES, Johnson B, Fitzhugh S, Gibson B, Butts CT. Warning tweets: serial transmission of messages during the warning phase of a disaster event. Inf Commun Soc 2014;17(6):765–87.
[45] Telford J, Cosgrave J, Houghton R. Joint evaluation of the international response to the Indian Ocean tsunami. Retrieved from http://www.alnap.org/resource/3535; 2006.
[46] Todd D, Todd H. Natural disaster response lessons from evaluations of the World Bank and others. Vol. 1. Retrieved from http://documents.worldbank.org/curated/en/2011/01/15512809/natural-disaster-response-lessons-evaluations-world-bank-others/; 2011.
[47] Varga I, Sano M, Torisawa K, Hashimoto C, Ohtake K, Kawai T, et al. Aid is out there: looking for help from Tweets during a large scale disaster. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Red Hook, NY: Curran Associates, Inc.; 2013. p. 1619–29.
[48] Veil SR, Buehner T, Palenchar MJ. A work-in-process literature review: incorporating social media in risk and crisis communication. J Conting Crisis Manag 2011;19(2):110–22.
[49] Vieweg S, Hughes AL, Starbird K, Palen L. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. Proceedings of the SIGCHI conference on human factors in computing systems. ACM; 2010, April. p. 1079–88.
[50] Villegas M, Paredes R. Image-text dataset generation for image annotation and retrieval. Prometeo 2009;14(2009):1–6.
[51] Westerman D, Spence PR, Van Der Heide B. A social network as information: the effect of system generated reports of connectedness on credibility on Twitter. Comput Hum Behav 2012;28(1):199–206.
[52] Woo H, Cho Y, Shim E, Lee K, Song G. Public trauma after the Sewol Ferry disaster: the role of social media in understanding the public mood. Int J Environ Res Public Health 2015;12(9):10974–83.
[53] Zheng, L., Guo, Y., Peeta, S., & Wu, B. (2019). Impacts of information from various sources on the evacuation decision-making process during no-notice evacuations in campus environment. J Transport Saf Secur, Advanced online publication.
[54] Zhang, Z., & Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:1805.07836.
[55] Zhao, W. (2017, August). Research on the deep learning of the small sample data based on transfer learning. AIP Conf Proc (vol. 1864, No. 1, p. 020018). AIP Publishing.