DisasterNet: Evaluating the Performance of Transfer Learning to Classify Hurricane-Related Images Posted on Twitter

Matthew Johnson
The University of Texas at Austin
mjohnson082396@gmail.com

Dhiraj Murthy
The University of Texas at Austin
dhiraj.murthy@austin.utexas.edu

Brett W. Robertson
The University of Texas at Austin
Brett.robertson@utexas.edu

W. Roth Smith
Illinois State University
wrsmit1@ilstu.edu

Keri. K. Stephens
The University of Texas at Austin
keri.stephens@austin.utexas.edu

Abstract

Social media platforms are increasingly used during disasters. In the U.S., victims consider these platforms to be reliable news sources and they believe first responders will see what they publicly post [1,2]. While having ways to request help during disasters might save lives, this information is difficult to find because non-relevant content on social media completely overshadows content reflective of who needs help. To resolve this issue, we develop a framework for classifying hurricane-related images that have been human-annotated. Our transfer learning framework classifies each image using the VGG-16 convolutional neural network and multi-layer perceptron classifiers according to the urgency, relevance, and time period, in addition to the presence of damage and relief motifs [3]. We find that our framework not only successfully functions as an accurate method for hurricane-related image classification, but also that real-time classification of social media images using a small training set is possible.

1. Introduction

Given that 9-1-1 emergency systems have experienced overloading during recent disasters in the US, social media data provides opportunities during a crisis for first responders and humanitarian non-governmental organizations (NGOs) to identify individuals who may need to evacuate or are seeking help. However, as previous work has highlighted, searches of social media during natural disasters have high frequencies of non-relevant content both for text [4] and images [5]. Though there is interest in studying images posted during hurricane events, much of the work has been focused around identifying the authenticity of images [6]. While there is some work seeking to identify hurricane-related images on social media through machine-learned methods, these approaches still demonstrate that 'modeling the relevancy of certain image content is another core challenge that needs to be addressed more rigorously' [7]. The main contribution of this study is to address this core challenge and evaluate whether we can obtain high performance classifiers with a limited number of human-annotated images. Ultimately, we find that we can use just 1,128 human-annotated images and transfer learning with the VGG-16 convolutional neural network to classify hurricane-related images for a range of attributes. These findings will enable first responders and humanitarian NGOs to identify images posted by those needing help using small training sets.

2. Related work

2.1. Images on social media

Though the posting of images and video during natural disasters has become an important part of how these crises are socially experienced and understood [3,7,8] studies of visual content during disasters remain heavily overshadowed by textual analyses. Indeed, there is a dearth of work exploring images posted on social media during disasters. Early work by Gupta and colleagues [6] 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. [3] 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 [3] advice "to develop ways of tackling these obstacles" for future crises that are socially experienced on Twitter.

2.2. Machine learning in disasters

Much work has been done regarding classifying Twitter data as relevant or non-relevant [9] ('signal' or 'noise'). For example, the Artificial Intelligence for Disaster Response (AIDR) system performs automatic classification. Though it is designed for tweets, it is specifically a text-only system. AIDR uses machine learning methods that are trained on human coded 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 [10]. Tweedr is another machine-based pipeline that uses tweets to extract actionable information for disaster relief workers [11]; this tool uses classification, clustering, and extraction.

Lagerstrom et al. [12] explored image classification using images from a bush fire in the Australian state of New South Wales. They used 6,214 images from tweets to classify images into fire and not fire-related classes, achieving an accuracy of 86%. While much of the past research with AIDR and Tweedr used text-based data for classification, recent work [9] studying the use of social media used during Hurricane Harvey employs supervised learning methods, specifically with images. O'Neal et al.'s [9] study, however, was not focused on Twitter, but instead evaluated the use of supervised learning based on samples of private social media data. While their work did not evaluate deep learning methods and did not use training sets developed from noisy, public social media platforms, 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).

One of the few studies to employ deep learning methods with disaster-related images is Nguyen et al.'s [7] study of two earthquakes, a typhoon, and a hurricane using data obtained from the AIDR platform. They used human annotations from AIDR and Crowdflower, a crowd-sourcing platform. For the four cases they studied, they assessed whether images reflected severe, mild, or no damage. Their work highlights the challenges of working with social media images – including redundant and irrelevant images since their overall results, through promising, reflected an F1 score of 0.67, leaving room for improvement.

3. Dataset

To collect data, we used Twitter's streaming API. Specifically, we used the API to collect data from the 'Spritzer' stream, a free data pipeline that allows researchers to collect 1% of all tweets at random, selected based on the time tweets are posted [13]. We studied tweets from August 17, 2017 to September 17, 2017 and collected all Hurricane Harvey-related tweets with the keywords: 'hurricane', 'harvey', 'hurricaneharvey', and 'harveyhouston'. From these tweets, we extracted all the media-related links to retrieve images. Duplicate images were removed by computing an MD5 checksum for each image. The resulting total number of images was 23,692. After duplicates and empty images were removed, 17,483 images remained.

To develop the training dataset for our study, we randomly sampled 1,128 images (approximately 6.45% of all images collected) and human-coded these images using a rubric with categories drawn from existing research. First, we coded time period, which was defined as pre-storm (August 17, 2017 to August 25, 2017; 124 images), landfall (August 26, 2017 to September 1, 2017; 735 images), and Harvey's aftermath/immediate cleanup (September 2, 2017 to September 17, 2017; 269 images). Urgency was the second category. Saldana [14] suggests that identifying the level of importance of a social media post by adding a magnitude rating to the coding scheme is vital. Therefore, images were rated (4 = highly urgent, 3 = moderately urgent, 2 = somewhat urgent, 1 = not urgent, 0 = spam/unrelated to Hurricane Harvey), similar to Iakovou and Douligeris' [15] recommendations on severity of a hurricane. The type of image in a disaster was drawn from Paul's [16] work on images posted to Twitter during the 2010–11 Queensland floods. Paul [16] argues that the major themes for tweets immediately following a disaster include requests, reports and reactions. Finally, the notion of an image motif was drawn from Murthy et al. [3] who stress that the reoccurring patterns of images posted during disasters allows viewers to understand the social experience of disasters. These motif categories included: ad, animals, damage, drink, food, gear, macro, outside, people, relief, and 'other'. In terms of motif classification, Paul [16] suggests these concepts are theoretically meaningful in disasters,

particularly relief and damage. For the purpose of this study, we choose to use Paul's [16] rationale to specifically investigate the 'damage' and 'relief' themes. Despite the comparatively small size of our dataset, recent efforts suggest that high-performance machine learning models can be constructed even with a dataset with limited samples [17].

4. Deep learning pipeline

In this section, we describe the methodology used to classify images by time period, relevancy, urgency, and the presence of relief and damage themes. A high-level overview of our framework is illustrated in Figure 1.

The first stage of our pipeline leverages transfer learning to extract features from the sample images. Transfer learning refers to applying knowledge gained by solving a prior problem to a new, but related problem. The effectiveness of deep learning methods, like Convolutional Neural Networks, is often limited by the size of the training set [18]; transfer learning offers the benefits of these deep learning methods without 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.

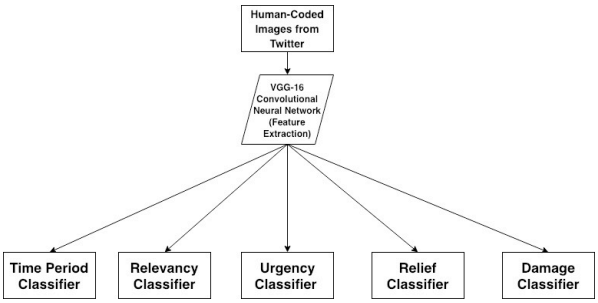

Figure 1. High-level overview of image classification pipeline

Instead of feeding raw images into models, images can be fed into convolutional neural networks to reduce their dimensionality. 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 and shadows, that represent a condensed version of 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 [19].

As part of the ImageNet Challenge 2014, Simonyan and colleagues [20] leveraged many small convolutional filters to create a highly accurate convolutional neural network that performed best at classifying images into 1000 categories. For the purpose of this study, we use their model, VGG-16, as our baseline architecture for the image classifiers we construct. Specifically, we collected the output from the second-to-last layer of VGG-16 before classification and treat this output as a "feature vector" constituting a low-dimensional representation of each image.

After extracting the feature vectors for each model, five multi-layer perceptron networks were constructed to classify images by time period, urgency, relevance, and the presence of 'damage' and 'relief' themes. We employed nearly identical model architectures for the binary classifiers ('relief' and 'damage' classifiers) and the multiclass classifiers (time period, relevancy, urgency), with the exception of the output layers.

For each model, the image and label pairs were independently and 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 80% of the total images and the validation set consisted of the remaining 20% of images. For each classifier, loss minimization was achieved using the Adam optimizer, which has been shown to converge faster compared to traditional methods like stochastic gradient descent [21].

Each binary classifier was optimized against the appropriate cross-entropy loss function,

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \quad (1)$$

where M is the total number of candidate categories, $y_{o,c}$ is a binary indicator if class label c is the correct classification for observation o, and $p_{o,c}$ is the predicted probability of observation o belonging to class c [22]. In addition, we recorded the accuracy of each classifier on both the training and validation set at each epoch of training. 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 a decrease in training loss is accompanied by a significant increase in the validation loss of the model over each epoch of training; this method prevents models from overfitting. Each epoch was trained in batches of 32. Moreover, each feed-forward network consisted of five dense layers, with each layer having twenty nodes.

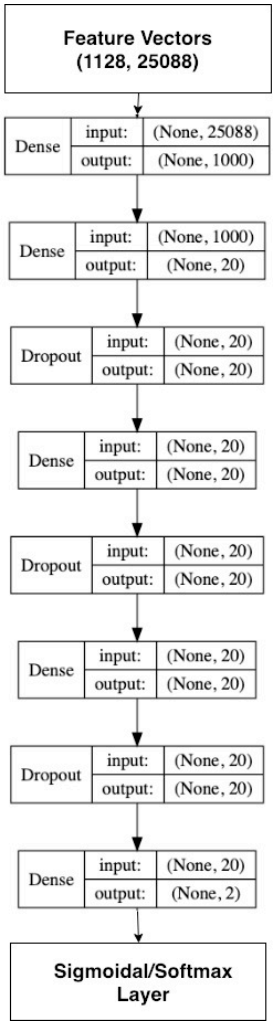

Figure 2. Generalized classifier model architecture

Each layer contained a rectified linear, or 'ReLu' activation function to promote sparse activation and efficient gradient propagation, which, in turn, promoted mathematical stability and computational efficiency [23; 24]. 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 was passed through either a sigmoidal or softmax activation layer so that the final output of each network represented a classification probability for each of the relevant classification categories. Figure 2 illustrates the generalized architecture used for all of the models constructed in this study.

5. Results

For each classifier, we collected the accuracy and categorical cross-entropy loss for the training and validation set. In addition, we evaluated the ability of our classifiers to discern between categories using a combination of micro-F1 score, macro-F1 score, and confusion matrices. In cases where classes were imbalanced, reporting predictive accuracy alone can be misleading since a classifier that simply classifies the most frequent class can be highly accurate but have very little predictive power [25]. In these scenarios, we examined the predictive power of the classifiers in the context of this class imbalance. Traditionally, the F1 score is computed as the harmonic mean between the recall and precision of a classifier; however, in the case where class imbalance exists, the micro-F1 score, which is the average of F1 scores across categories weighted by class occurrence, can be a stronger indicator of classifier performance [26]. For this reason, we report both the macro (unweighted) and micro (weighted) F1 scores of our classifiers.

Furthermore, we report the confusion matrix for each classifier as a means to visualize the predictive power of each model across categories. An entry at row *i* and column *j* of a confusion matrix represents the number of samples predicted as belonging to class *i* whose ground-truth classification is the class represented by *j*. The confusion matrices and F1 scores were computed on the validation set. Since each classifier's last layer outputs a vector of probabilities for each category, we took the classifier's prediction to be the category with the largest softmax output. The performance of our models is summarized in Table 1 below.

Classifier	Time Period	Urgency	Relevancy	Damage	Relief
Training Loss	0.6398	0.9514	0.3313	0.1441	0.128
Training Accuracy	0.7705	0.6098	0.9035	0.9957	0.9484
Validation Loss	0.7515	1.0841	0.5222	0.253	0.2335
Validation Accuracy	0.677	0.6195	0.8186	0.9023	0.9336
F1 Macro Score	0.3735	0.5847	0.752	0.6973	0.824
F1 Micro Score	0.5012	0.6157	0.811	0.4375	0.7568

Table 1. Summary of classifier performance

5.1. Time period classifier

The first classifier we constructed predicted the time period defined as pre-storm (August 17, 2017 to August 25, 2017; 124 images), landfall (August 26, 2017 to September 1, 2017; 735 images), and Harvey's

aftermath/immediate cleanup (September 2, 2017 to September 17, 2017; 269 images). The time period classifier ultimately reached a training accuracy of 0.7705, a training loss of 0.6398, a validation accuracy of 0.6770, and a validation loss of 0.7515 in 9 epochs. Furthermore, the time-period classifier recorded a macro-F1 score of 0.3735 and a micro-F1 score of 0.5012. Figure 3 shows a heat-mapped confusion matrix for the time-period classifier evaluated on the validation data.

The disparity between the macro-F1 score and micro-F1 score, coupled with the confusion matrix, illustrates the strengths and limitations of the time-period classifier. The classifier is notably adept at correctly identifying images labeled as 'landfall'; however, the classifier struggles to accurately identify images labeled as 'pre-storm' and 'cleanup'. This inaccuracy suggests that additional coded images may be required to help for our framework discern between time periods; it is likely the case that images posted throughout the storm share strong similarities regardless of time of posting.

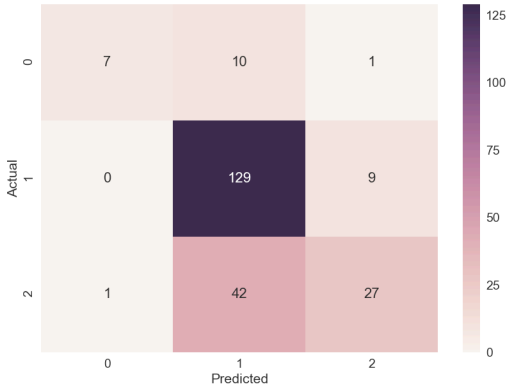

Figure 3. Confusion matrix for the time-period classifier

5.2. Relevancy classifier

The next classifier constructed predicted the 'relevancy' of each image defined as either irrelevant ('0', 152 images), relevant ('1', 736 images), or uncertain ('2', 239 images). The relevance classifier ultimately reached a training accuracy of 0.9035, a training loss of 0.3113, a validation accuracy of 0.8186, and a validation loss of 0.5222 in 18 epochs. Furthermore, the relevancy classifier recorded a macro-F1 score of 0.7520 and a micro-F1 score of 0.811. Figure 4 shows a heat-mapped confusion matrix for the relevance classifier evaluated on the validation data.

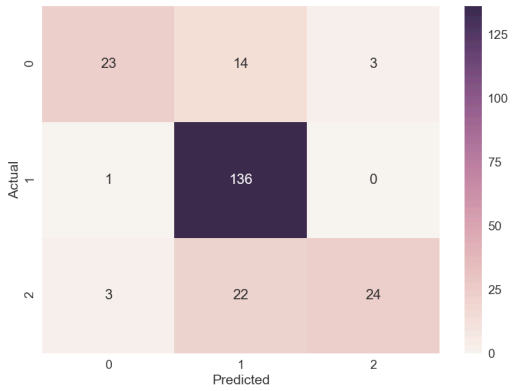

Figure 4. Confusion matrix for relevance classifier

Both the micro-F1 and macro-F1 scores recorded by the relevancy classifier surpass the results obtained by Nguyen et al. [7]. These results, combined with the confusion matrix (see Figure 4), support the successful performance of our transfer learning framework over existing approaches.

5.3. Urgency classifier

The urgency classifier was trained to predict the urgency depicted in an image into different levels. Here, images were ranked in terms of importance to Hurricane Harvey (4 = highly urgent [87 images], 3 = moderately urgent [181 images], 2 = somewhat urgent [352 images], 1 = not urgent [151 images], 0 = spam/unrelated to Hurricane Harvey [356 images]), similar to Iakovou and Douligeris' [15]) recommendations on severity of a hurricane. The urgency classifier ultimately reached a training accuracy of 0.6098, a training loss of 0.9514, a maximum validation accuracy of 0.6195, and a validation loss of 1.0841 in 18 epochs. Furthermore, the urgency classifier recorded a macro-F1 score of 0.5841 and a micro-F1 score of 0.6157. Figure 5 shows a heat-mapped confusion matrix for the urgency classifier evaluated on the validation data.

Both the micro-F1 and macro-F1 scores approach the benchmarks established by Nguyen et al. [7], despite the inclusion of additional levels of urgency. Of the 124 images in the validation set that were at least 'somewhat urgent', our framework correctly identified 86% as at least 'somewhat urgent'. This suggests that our framework, even with a limited dataset, can be used to filter images by urgency from the images alone with a degree of accuracy and discretionary power at least as strong as existing published work [7].

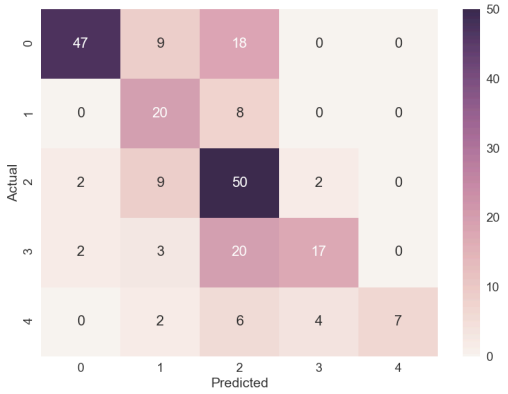

Figure 5. Confusion matrix for urgency classifier

Furthermore, our framework provides a method of classifying images into 'levels' of urgency, which could allow first responders and agencies to prioritize aid to areas where a poster's images are identified as 'moderately urgent' or 'highly urgent'.

5.4. Relief classifier

The relief classifier was trained to identify which images contained the 'image motif' of relief identified by Murthy et al. [3], who stressed that reoccurring patterns of images posted during disasters allow viewers to understand the social experience of disasters. The 'relief' motif consisted of 'images depicting relief efforts and relief campaigns'. In this study, images were marked as either containing the 'relief' theme ('1'; 82 images) or not ('0'; 1880 images). Ultimately, the relief classifier reached a training accuracy of 0.9484, a training loss of .1280, a validation accuracy of 0.9336, and a validation loss of 0.2335. Moreover, the relief classifier held a macro-F1 score of 0.8240 and a micro-F1 score of 0.7568. Figure 6 shows a heat-mapped confusion matrix for the relief classifier evaluated on the validation data.

While the relief classifier demonstrated accuracy and F1 scores commensurate with prior work, the classifier showed little ability to discern between images that contained the relief motif. Of the 17 images in the validation set containing the relief motif, only 42% were accurately classified as relief-related images. It is likely that more relief-related images are needed to improve classifier performance, despite the flexible methodology we developed in this study.

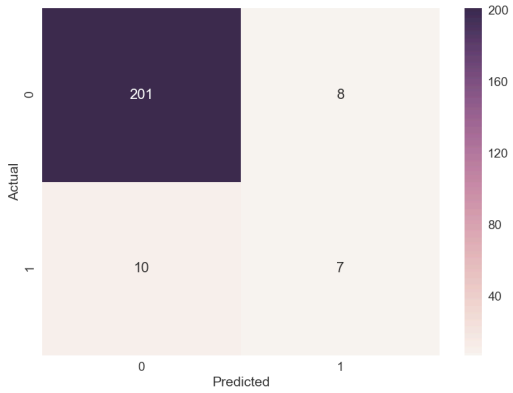

Figure 6. Confusion matrix for relief classifier

5.5. Damage Classifier

Similar to the relief classifier, the damage classifier was trained to identify which images contained the 'image motif' of 'damage defined by Murthy et al. [3], as 'images depicting storm-related damage to the built environment or otherwise.' In this study, images were marked as either containing the 'damage theme ('1'; 295 images) or not ('0'; 1867 images). Ultimately, the damage classifier reached a training accuracy of 0.9557, a training loss of .1441, a validation accuracy of 0.9027, and a validation loss of 0.2530. Moreover, the relief classifier held a macro-F1 score of 0.6973 and a micro-F1 score of 0.4375. Figure 7 shows a heat-mapped confusion matrix for the damage classifier evaluated on the validation data.

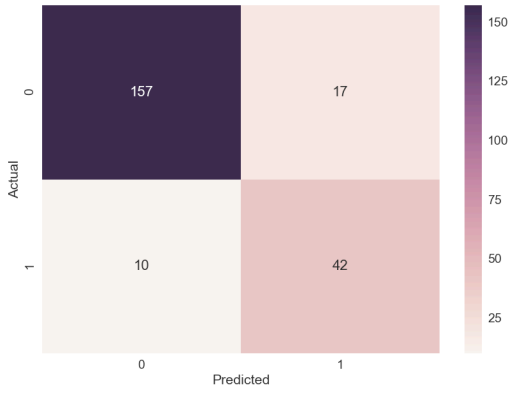

Figure 7. Confusion matrix for damage classifier

At first glance, the damage classifier outperformed the relief classifier under the framework developed in this study: Of the 52 images in the validation set containing the damage image motif, 82% were

classified correctly. This result could be because there were significantly more images in our dataset that were tagged with the damage motif than the relief motif. These results suggest that our framework could be used to filter images related to damage with a limited number of candidate images and without additional metadata.

6. Conclusion

Existing approaches [7] using deep learning methods indicate state-of-the-art baselines with F1 scores ranging from 0.60-0.70. Our results, particularly with the relevancy, urgency, and damage classifiers, provide evidence that high-performance, robust classifiers (with accuracy rates of 82%, 62%, and 91%, and F1-macro scores of 0.7520, 0.5847, and 0.6973 respectively) can be obtained, even with a limited number of human-annotated images. The improvement we achieved over existing baselines is due in part to: (1) the use of highly trained human annotators with field experience in disaster contexts as opposed to crowd-sourced platforms, and (2) an improved transfer-learning pipeline using the VGG-16 CNN and multi-layer perceptrons, a process which requires significantly less training data than existing approaches but still achieves comparable accuracy and F1 scores.

Our study provides evidence that if transfer learning is used to build models to filter images by urgency, relevancy, and damage, with a limited amount of training data, custom models do not have to be built. Rather, by implementing our framework into a data pipeline, stakeholders—including first responders and humanitarian NGOs—can label a small volume of social media images. During future hurricane events, this will not only save time for those tasked with emergency response activities but makes possible near real-time analysis of large numbers of images gathered from social media.

6.1. Limitations and future directions

Several of the classifiers constructed were highly accurate (reaching validation accuracies of no less than .616) and matched or exceeded the benchmarks for F1 scores established by Nguyen et al. [7]. In spite of this success, our results with the relief and damage classifiers suggest that our methodology could be improved in cases of extreme class imbalance, even though our framework was quite successful with a small dataset. Because of the high variability inherent in images posted on Twitter, the true proportion of images labeled with these themes is low; this poses challenges to researchers and stakeholders attempting to build and train autonomous systems to identify these informative images before, during, and after a disaster. To remedy this, we are currently investigating purchasing larger datasets directly from Twitter to increase dataset size in future studies. In addition, future work will include an investigation into different baseline models (other than VGG-16), and an evaluation of model architectures other than the multi-layer perceptron to classify images from feature vectors.

7. Acknowledgments

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 recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. References

[1] W.R. Smith, K.K. Stephens, B.W. Robertson, J. Li, and D. Murthy, "Social media in citizen-led disaster response: Rescuer roles, coordination challenges, and untapped potential", Proceedings of the 15th International Information Systems for Crisis Response and Management (ISCRAM) Conference. Rochester, NY, 2018.

[2] K.K. Stephens, J. Li, B.W. Robertson, W.R. Smith, and D. Murthy, "Citizens communicating health information: Urging others in their community to seek help during a flood", Proceedings of the 15th International Information Systems for Crisis Response and Management (ISCRAM) Conference. Rochester, NY, 2018.

[3] D. Murthy, A. Gross, and M. McGarry, "Visual social media and big data: Interpreting Instagram images posted on Twitter", Digital Culture & Society, 2016, pp. 113-134.

[4] A. Cobo, D. Parra, and J. Navón, "Identifying relevant messages in a Twitter-based citizen channel for natural disaster situations", Proceedings of the 24th International Conference on World Wide Web. ACM Press, New York, NY, 2015.

[5] R. Peters, and J. Porto de Albuquerque, "Investigating images as indicators for relevant social media messages in disaster management", Proceedings of the 12th International Information Systems for Crisis Response and Management (ISCRAM). Kristiansand, Norway, 2015.

[6] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, "Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy", Proceedings of the 22nd

International Conference on World Wide Web. New York, NY, 2015.

[7] D. Tien Nguyen, F. Alam, F. Ofli, and M. Imran, "Automatic image filtering on social networks using deep learning and perceptual hashing during crises", Proceedings of the 14th International Information Systems for Crisis Response and Management (ISCRAM) Conference. Albi, France, 2017.

[8] K.K. Stephens, B.W. Robertson, & D. Murthy, "Throw me a lifeline: Articulating mobile social network dispersion and the social construction of risk in rescue communication", Mobile Media & Communication, 2019, Advanced online publication.

[9] A. O'Neal, B. Rodgers, J. Segler, D. Murthy, N. Lakuduva, M. Johnson, and K.K. Stephens, "Training an emergency-response image classifier on signal data", Proceedings 17th IEEE International Conference on Machine Learning and Applications (ICMLA). Orlando, FL.

[10] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: Artificial intelligence for disaster response", Proceedings of the 23rd International Conference on World Wide Web. New York, NY, 2014.

[11] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta, "Tweedr: Mining twitter to inform disaster response", Proceedings of the 11th International Information Systems for Crisis Response and Management (ISCRAM) Conference. University Park, PA, 2014.

[12] R. Lagerstrom, Y. Arzhaeva, P. Szul, O. Obst, R. Power, B. Robinson, and T. Bednarz, "Image classification to support emergency situation awareness", Frontiers in Robotics and AI, 2016, pp. 1-11.

[13] D. Kergl, R. Roedler, and S. Seeber, "On the endogenesis of Twitter's Spritzer and Gardenhose sample streams", Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Beijing, China, 2014.

[14] Saldana, J.M., The Coding Manual for Qualitative Researchers (2nd. ed.). Sage, London, UK, 2012.

[15] E. Iakovou, and C. Douligeris, "An information management system for the emergency management of hurricane disasters", International Journal of Risk Assessment and Management, 2001, pp. 243-262.

[16] A. Paul, "Identifying relevant information for emergency services from twitter in response to natural disasters", Retrieved from https://eprints.qut.edu.au/89220/1/Avijit_Paul_Thesis.pdf, 2015.

[17] M. Hussain, J.J. Bird, and D.R. Faria, "A study on CNN transfer learning for image classification", UK Workshop on Computational Intelligence. Nottingham, UK, 2018.

[18] H. Hosseini, B. Xiao, M. Jaiswal, and R. Poovendran, "On the limitation of convolutional neural networks in recognizing negative images", 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 2017.

[19] L. Hertel, E. Barth, T. Käster, and T. Martinetz, "Deep convolutional neural networks as generic feature extractors", International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 2015.

[20] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv:1409.1556, 2014.

[21] D.P. Kingma, and J. Ba, "Adam: A method for stochastic optimization", arXiv:1412.6980, 2014.

[22] Z. Zhan, and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels", 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montreal, Canada, 2018.

[23] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks", Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL, 2011.

[24] V. Nair, and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines", Proceedings of the 27th International Conference on Machine Learning (ICML-10). Haifa, Israel, 2010.

[25] J. Akosa, "Predictive accuracy: A misleading performance measure for highly imbalanced data", Proceedings of the SAS Global Forum. Orlando, FL, 2017.

[26] M. Sokolova, and G. Lapalme, "A systematic analysis of performance measures for classification tasks", Information Processing & Management, 2009, pp. 427-437