Training an Emergency-Response Image Classifier on Signal Data

Aubrey O'Neal,[1] Benjamin Rodgers,[2] Justin Segler,[2] Dhiraj Murthy,[1]
Nandhini Lakuduva,[2] Matthew Johnson,[3] and Keri Stephens [4]
Departments of Journalism,[1] Computer Science,[2] Electrical and Computer Engineering,[3] and Communication Studies [4]
University of Texas at Austin
Email: aoneal@utexas.edu, rodgers.benjamin@utexas.edu, justin@justinsegler.com, dhiraj.murthy@austin.utexas.edu,
nlakuduva@utexas.edu, mjohnson082396@utexas.edu, keri.stephens@austin.utexas.edu

***Abstract*— The increasing popularity of multimedia messages shared through public or private social media spills into diverse information dissemination contexts. To date, public social media has been explored as a potential alert system during natural disasters, but high levels of noise (i.e. non-relevant content) present challenges in both understanding social experiences of a disaster and in facilitating disaster recovery. This study builds on current research by uniquely using social media data, collected in the field through qualitative interviews, to create a supervised machine learning model. Collected data represents rescuers and rescuees during the 2017 Hurricane Harvey. Preliminary findings indicate a 99% accuracy in classifying data between signal and noise for signal-to-noise ratios (SNR) of 1:1, 1:2, 1:4, and 1:8. We also find 99% accuracy in classification between respondent types (volunteer rescuer, official rescuer, and rescuee). We furthermore compare human and machine coded attributes, finding that Google Vision API is a more reliable source of detecting attributes for the training set.**

Keywords- Emergency response, crisis communication, image classification, machine learning

I. Introduction

During Hurricane Harvey, the US emergency telephone hotline, 9-1-1, was overwhelmed, causing residents to turn to social media [1]-[2]. Current research has attempted to create machine learning models based on data from public social media, in order to assess need for emergency response during natural disasters [3]-[4]. The high level of noise and difficulty in discerning the users location prompts this study to explore a different solution. We hypothesized that, by collecting private social media messages from recruited participants, we could construct a supervised machine learning model from a training set with high signal. Additionally, this project leverages computer vision in order to bolster the training of our models. The modular design of this project enables many potential applications.

Crisis informatics literature has focused its methodology on collecting vast amounts of data from public social media APIs (particularly Twitter) with inclusion criteria based on a combination of keywords, date ranges, and other attributes.

Much of this work became popular due to the highly API- accessible 1% Spritzer stream on Twitter, which allows anyone to collect up to 1% of global tweets. This has been referred to some by some as a socio-scope [5], providing insights into any event, however large or small. There is, of course, value in such blanket approaches and the literature has gained much in terms of understanding disasters both from the point of view of victims as well as first responders. However, there is a dearth of work that has involved actually deploying teams in the field during a disaster to assess how social media was used, in addition to collecting data more indicative of a disaster experience that are circulated on private social media networks, particularly Facebook and Nextdoor. Though current state-of-the-art methods are able to classify the relevancy of content to a disaster, these rates have much room for improvement.

Current work in crisis informatics and machine learning focuses on the challenges of high-volume, high-velocity data scraped from social media outlets. Scholars agree that one of the biggest challenges lies in differentiating noise from signal in an accurate and timely manner [6]-[7]. As inexpensive image classification APIs such as Google Vision continue to improve, machine image analysis is becoming an increasingly viable option for research [7]. Reference [8] had a high overall level of success (AUC: 0.98, Precision: 0.99 Recall: 0.97, and F1: 0.98) in creating a pipeline to filter irrelevant and redundant imagery through a combination of image classification and human curation. However, this model does not train on field-elicited data and is not scalable due to use of human annotators. Few crisis informatics studies have trained on highly curated non- public social media data.

II. Methods

A. Overview

Using attributes detected in images via the Google Vision API, we constructed a supervised machine learning model. Fig. 1. illustrates our methodological process from gathering images to extracting attributes and performing a frequency analysis on most common attributes, to training our model. We repeated the same process twice in pretests, comparing Google Vision API to human coders at the image attribute detection phase. Training data consisted of private images gathered through fieldwork over several months directly following Hurricane Harvey (August 17, 2017 September 3, 2017). The machine learning prediction model was trained on this signal-based dataset and success for phase 1 of the

project was measured by the models ability to classify images between signal and noise data; phase 2 of the project is to test categorization of media into one of three categories: official rescuer, volunteer rescuer, and rescuee. The model was tested at scale with 37,500 noisy images, randomly extracted from the Twitter 1% spritzer stream from May 1, 2017 -January 20, 2018. The training of the model was modified as signal-to-noise ratio (SNR) was increased and as new classifiers were tested. Methods such as cross validation and stacking were used to limit bias and maximize classification accuracy.

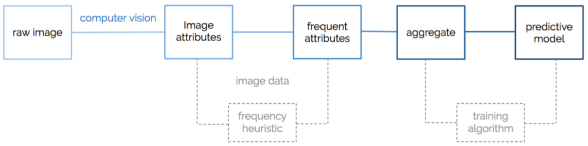

Fig. 1. Process Diagram

B. Data Collection

The data collection approach for the study was developed following classic methods of identifying the relevant stages of a disaster: pre-impact, impact, and recovery [9]-[10]. Previous work studying social media and disasters has generally collected content at all of these stages from platforms such as Twitter [11], Facebook public groups [12], and Instagram [13]. Reference [13] work indicates high volumes in the first two stages, particularly affected by news reports and celebrity mentions of disasters on social media. It is also for this reason that searching for a disaster by keywords (in the case of Hurricane Harvey, by #Harvey) tend to have high levels of noisy data versus data that is more representative of individuals experiencing the disaster (i.e. high signal data). For this reason, data collection involved multiple trips to affected areas of Houston to identify official rescuers, volunteer rescuers, and those affected by the disaster (whom we refer to as rescuees).

During each of these field site visits, a member of the research team interviewed individuals following an approved Institutional Review Board (IRB) protocol, which involved the method of photo elicitation interview (PEI) [14], in which respondents were asked to contribute their social media activity to the research team. These included photos and videos taken during Hurricane Harvey as well as posted and received textual content as part of their rescue experience or by those actually conducting rescues.

When consented by the respondent, comments were also captured in screenshots and shared with the research team. As interviews took place between 15 minutes to more than an hour, multiple opportunities were available to collect these types of "private" data which would be inaccessible to those acquiring data from public APIs. Data collected in the form of screenshots were deposited in a central secure repository. Fieldworkers consisted of a team of trained graduate students and faculty at a public university, from multiple disciplines. Overall, images, text, and media screenshots were gathered from fieldwork, though our classifier utilizes only images (see Table 1). The majority of the content the research team collected were from Facebook.

TABLE I
SIGNAL DATA

	Signal Data Source			
Data Type	Volunteer Rescuer	Official Rescuer	Rescuee	Total
Image	158	36	248	442

III. Attribute Detection: Google Vision API

For the purposes of this project, both signal and noise images were processed by Google Vision through a built streamlined process for attribute detection. Google Vision identified attributes in each of the images and returned a structured JSON list (e.g. water, flood, and boat). For all media passed through Google Vision, personal information such as names and profile pictures were redacted.

A. Frequency Analysis

For images processed by Google Vision, detected attributes were also assessed through a frequency analysis. We aggregated all the attributes identified from computer vision into a single attribute set. We then calculated the frequency at which each attribute occurred in the image set. The threshold value determined the minimum frequency of included computer vision attributes. For example, a threshold of 0.00 included any attribute identified by computer vision whereas a threshold of 0.10 included only attributes which were identified in at least 10% of the images and excluded any other infrequent attributes.

When training the machine learning model, we varied the minimum frequency threshold required, which limited the attributes utilized in training. This was done to assess whether more frequent and related data increased success, speaking to our larger research goal of exploring a signal-centric methodology. Attributes above a set threshold were aggregated, and this set of attributes became the features to be trained upon.

B. Human Attribute Detection Versus Google Vision API

In addition to using Google Visions machine coding to gather attributes, traditional human coding techniques were explored in pre-testing. To do so, we followed the same process described in Fig 1. Studies have reported success in using human coding, though success was not directly compared to computer vision attribute detection [8]. The goal of comparing the two in pre-tests was to ask the methodological question, "Can a machine identify features in an image as well as a human?"

Similar to the functions of Google Vision API, the human codebook allowed coders to record manifest attributes found in the media (i.e. car, house, and water), in addition to latent attributes (i.e. phenomenon, disaster), without any restrictions such as a predefined dictionary. We found that the

human coders provided fewer attributes, misinterpreted some attributes, and were potentially biased because they were aware they were coding images related to Hurricane Harvey. Based on results of a pre-test comparing the models accuracy when trained on machine versus human attributes, we chose to rely on Google Vision attributes. Fig. 2 reveals that the classifier trained on human-coded yielded low accuracy results in comparison to attributes from Google Vision.

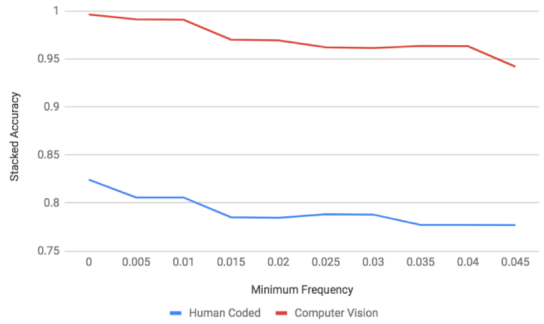

Fig. 2. Minimum frequency vs. stacked accuracy for classifiers trained on attributes detected by human coders and by computer vision.

IV. Training And Testing

The next phase of the project involved the development of a classifier, whose aim was to classify content at scale from noisy data that was relevant to Hurricane Harvey. In other words, our methodology goes from signal-to-noise, rather than from noise-to-signal. By starting out with a high quality, fieldwork-elicited training data set, our hope was to develop a classifier with very high accuracy.

After collecting the data and creating records for each image using the attributes detected through Google Vision, the data were shuffled so that the inherent class imbalance of the data did not affect the outcome. By randomizing the ordering, we reduced the possibility that an unrealistic sampling of the data was used during training and testing phases. We created base classification models for support vector machine (SVM), Gaussian naive Bayes (GNB), multinomial naive Bayes (MNB), Bernoulli naive Bayes (BNB), k nearest neighbor (KNN), decision tree (DT), stochastic gradient descent (SGD), and multilayer perceptron (MLP). For each classifier, we fit the model using a 5-fold cross validation. For SVM, KNN, and SGD, we used scikit-learns gridsearchcv functionality to tune our hyperparameters which uses nested cross validation over each combination of hyperparameters [16].

The results from these 8 base classifiers were then used to create a stacked classifier. The predictions from the base classifiers replaced the original features used in the previous iteration and the stacked classifier was trained using only the base predictions. No other features from the original feature set were included in the stacked classifier. Ensemble methods and a voting classifier were also tested, but proved less accurate than the stacked classifier, though higher than any base classifier. The ensemble classifier was created by randomly sampling the dataset with replacement. This process was repeated until we had 5 unique samples. Each sample was trained using the same machine learning algorithm and the predictions from each base model were then used to vote for the final prediction. The voting classifier was created using the same base classifiers as the stacking classifier, but followed the same majority rule as the ensemble methods for determining the ultimate prediction.

We experimented with naive Bayes, KNN, SVM, and MLP for the stacked classifier, and found that in most cases, the algorithms performed similarly. As demonstrated in Fig. 3., SVM and MLP were marginally better than KNN and naive Bayes was slightly worse. The models F1 scores reinforce these results as seen in Fig. 3. For further testing, the 8 base classifiers were stacked into an optimally tuned SVM classifier.

We tested accuracy and F1 score of models by manipulating the level of signal data in addition to threshold of attribute frequency. To gather samples, tweets were randomly selected from the Twitter repository. Samples of sizes 500, 1000, 2000, and 4000, each with 5 SNR permutations, were tested for a total of 20 different models. Attribute frequency threshold values were tested from 0.0-0.045 with 0.005 steps. Additionally, each permutation set was tested on every threshold. For each combination, accuracy and F1 scores were recorded.

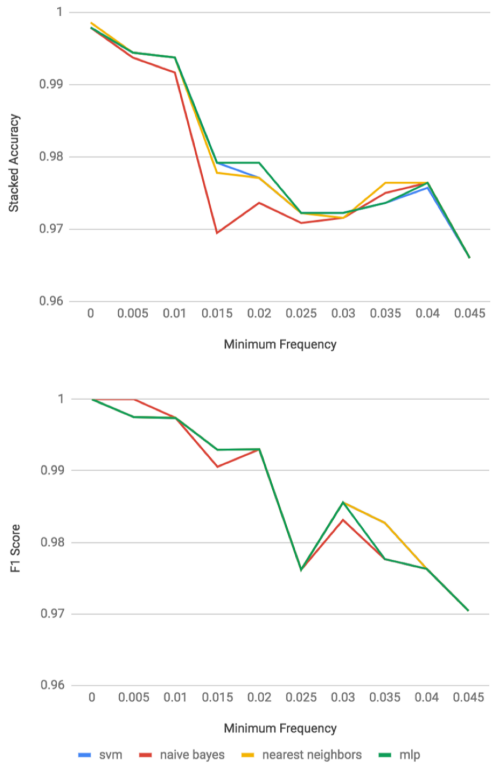

Fig. 3. Threshold vs accuracy for stacked classifier algorithm comparison. Attributes identified by Google Vision that occur with a frequency below the threshold value were not considered during machine learning steps.

V. Results

This model differentiates between relevant signal data and spurious noise data. The performance of the SNR classifier is assessed based on stacked accuracy and 8 base classifiers (see Fig. 4.), and on an F1 score (see Fig. 5.). An F1 score measures the performance of a classifier, taking into account the model's accuracy in classifying both positive cases and negative cases [15].

The classification accuracy achieved is high for all signal-to-noise data sets, though notably stacked accuracy falls as SNR approaches a 1:1 ratio. To back the high accuracy achieved in initial testing, we performed a second experiment to visually represent our data. The eight features gathered from the base classifiers were projected onto a 2-dimensional scatter plot using singular value decomposition. The resulting graph showed a clear linear separation between signal and noise data points, reinforcing the high accuracy of an SVM classification model.

Preliminary results of an additional model trained to classify signal data between image types indicate a 99% stacked accuracy for a threshold between 0.00 and 0.005. This model differentiates between images representing rescuees people who were in need of rescue at the time of Hurricane Harvey and rescuers people who took part in rescue efforts at the time of Hurricane Harvey.

A. Misclassification Evaluation

We see an accuracy improvement as we train a model with more noise compared to signal data. This may be due to the fact that we are misclassifying signal results at a similar rate per SNR permutation, but because there are more results, the cost of misclassification is smaller for larger noise ratios. To evaluate this theory, we built a precision recall confusion matrix [16] as summarized in Table 2 and 3. Ideally, the count for true positives (TP) and true negatives (TN) should be high and false positives (FP) and false negatives (FN)

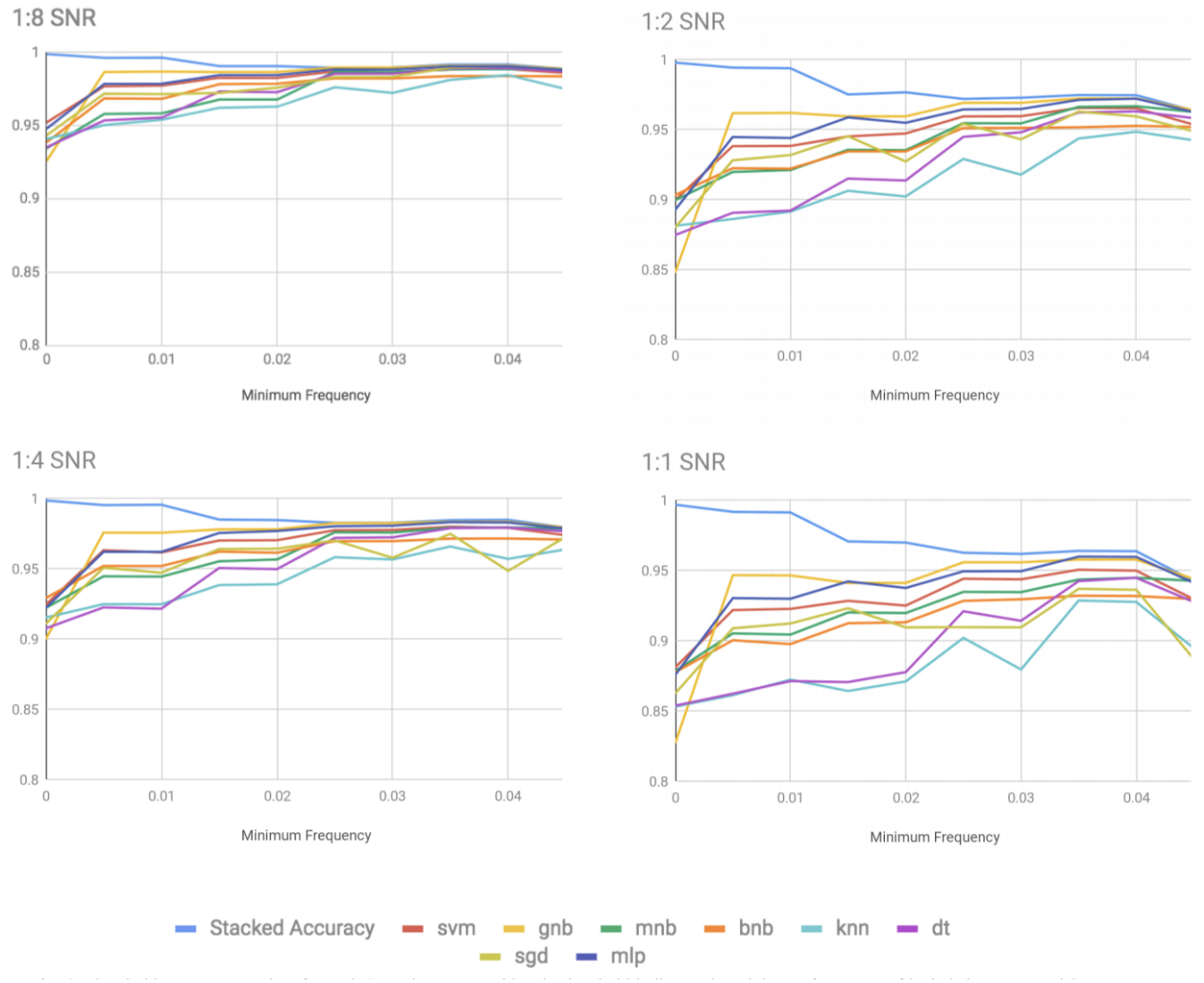

Fig. 4. Threshold vs. accuracy plots for each SNR data set considered. Threshold indicates the minimum frequency of included computer vision attributes. Labels identified by Google Vision that do not exist in at least the threshold value were not considered during machine learning steps.

should be low relative to TP and TN. Misclassification of signal as noise would be a false negative. As SNR increases, we see proportionally fewer FN, as the number of TN increases. Increasing the noise sample size doesnt dispose the model to classify all data points as noise, as it is still able to differentiate signal from noise and at a higher confidence.

$$Precision = \frac{|TP|}{|TP + FP|} \qquad Recall = \frac{|TP|}{|TP + FN|}$$

Precision and recall measure the accuracy of classifiers from a different point of view. (Unweighted) precision is defined as the fraction of records that actually are of class C, out of records predicted to be of class C. That is, given a positive prediction from the classifier, we ask how likely is it to be correct. In this case, as noise increases, we also see an increase in records that are correctly predicted to be noise. Recall is defined as the fraction of correct predictions of class C over all points in class C and answers the question: given a positive example, will the classifier detect it?

An ideal classifier sees that as precision increases, recall increases as well. Average precision measures this concept by calculating the weighted mean of precisions achieved at each SNR, with the increase in recall from the previous SNR used as the weight. The high trend of average precision indicates there is a direct rather than inverse relationship between recall and precision as SNR approaches 1:8. From the confusion matrices in Table 3, we see that the total number of misclassifications, the sum of FN and FP, remains constant while the number of correct predictions, the sum of TP and TN, increases. Thus, there are proportionally fewer misclassifications and a higher total accuracy as SNR approaches 1:8. However, the accuracy of signal classification remained steady, supported by the constant precision across all SNR permutations.

TABLE II
CONFUSION MATRIX KEY

		Predicted Class	
		Signal	Noise
Actual Class	Signal	F++ (TP)	F+- (FN)
	Noise	F-+ (FP)	F– (TN)

TABLE III
PRECISION AND RECALL EVALUATION

Sample Size	Unweighted Precision (+)	Average Precision	Average Confusion Matrix (n=5)
500	0.9982	0.9998	[[115.6 0.6] [0.2 119.6]]
1000	0.9947	0.9985	[[113.2 0.6] [0.6 246.6]]
2000	0.9964	0.9996	[[112.8 0.4] [0.4 497.2]]
4000	0.9941	0.9988	[[102.6 1.2] [0.6 1007]]

B. Threshold

Threshold indicates the minimum frequency of included computer vision attributes. The threshold value was explored in order determine whether excluding attributes specific to a small fraction of signal images would affect overall accuracy.

We hypothesized that a frequency threshold greater than 0.00 would yield a higher accuracy in image classification. However, the results showed a different trend. Between a threshold of 0.00 and 0.005, there was no significant difference; the best model considered all attributes, regardless of frequency of occurrence or human-coded relatedness. Fig. 5 compares the results from each frequency threshold across each SNR

As the threshold increases, accuracy of the model steadily declines. We concluded that having the largest number of attributes as possible for signal data collected from fieldwork yields the best results.

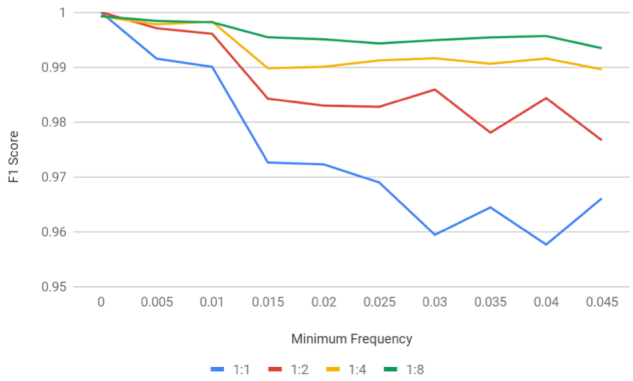

Fig. 5. Threshold vs. F1 Score for each SNR data set considered.

VI. CONCLUSION

In this paper we have implemented a pattern recognition machine for automatic classification of human roles in a natural disaster. A major contribution of the paper lies in pursuing hard-to-reach signal data, rather than noisy social media data, in addition to utilizing a stacked classifier modeling approach in the machine training. The results of extensive testing performed on multiple datasets of varying levels of noise data illustrates the robustness and potential advantages of this proposed approach.

There is still room for improvement in this classifier. By introducing different types of data and by testing it on noise data from other contexts, accuracy can be improved.

This is phase one of a crisis communication machine learning project. In ongoing research, we are integrating text data into the training model, with the goal of increasing context of the natural disaster data (Fig. 6). This will improve the models classification ability when faced with confounding social media imagery (i.e. lakes, rivers, and weather reports). We plan to conduct further noisy data tests using confounding social media imagery pulled from Hurricane Harvey-related tweets. Future work will also concern the

integration of this classifier with automated data collection directly from private and public social media streams.

The model developed in our lab has potential application as an alternative emergency response system, highlighting groups or individuals who are potential rescuers or who are in need of rescuing. There are many useful applications for integrating this project into existing social media platforms. For example, in times of disaster, bots could continuously surf these social media sites and scrape data to be passed to the model. Once passed to the model, content from social media sites can be flagged as either signal or noise, and if signal, rescuer or rescuee.

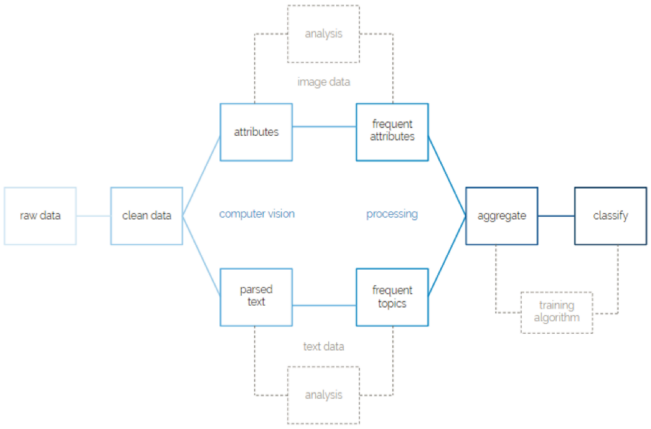

Fig. 6. Integrating imagery and text in the project work flow (See Fig. 1.).

ACKNOWLEDGMENT

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.

REFERENCES

[1] J. Cowan, When 911 failed them, desperate Harvey victims turned to social media for help, Dallas Morning News, August 27, 2017. [Online].

[2] M. Rhodan, 'Please Send Help.' Hurricane Harvey Victims Turn to Twitter and Facebook, Time, August 30, 2017. [Online].

[3] R. Lagerstrom, Y. Arzhaeva, P. Szul, O. Obst, R. Power, B. Robinson, and T. Bednarz, Image classification to support emergency situation awareness, Frontiers in Robot AI. 2016, doi:10.3389/frobt.2016.00054

[4] L. Palen, K.M. Anderson, G. Mark, J. Martin, D. Sicker, M. Palmer, and D. Grunwald, A vision for technology-mediated support for public participation & assistance in mass emergencies & disasters. Proceedings of the 2010 ACM-BCS visions of computer science conference. British Computer Society, 2010.

[5] Y. Mejova, I. Weber, and M.W. Macy, Twitter: a digital socioscope, Cambridge University Press, 2015.

[6] J. Qadir, A. Ali, R. ur Rasool, A. Zwitter, A. Sathiaseelan, and J. Crowcroft, Crisis analytics: big data-driven crisis response. Journal of International Humanitarian Action 1, no. 1 (2016): 12.

[7] Bonaretti, Dario, Effective Use of Twitter Data in Crisis Management: The Challenge of Harnessing Geospatial Data, unpublished.

[8] D.T. Nguyen, F. Alam, F. Ofli, and M. Imran, Automatic image filtering on social networks using deep learning and perceptual hashing during crises. arXiv preprint:1704.02602 (2017).

[9] E.L. Quarantelli, Enrico Louis. Disaster crisis management: A summary of research findings. Journal of management studies 25.4 (1988): 373- 385.

[10] S. Vieweg, A.L. Hughes, K. Starbird, L. Palen. 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.

[11] C.C. David, J.C. Ong, and E.F.T. Legara, Tweeting Supertyphoon Haiyan: Evolving functions of Twitter during and after a disaster event. PloS one 11.3 (2016): e0150190.

[12] Bird, Deanne, Megan Ling, and Katharine Haynes. "Flooding Facebook- the use of social media during the Queensland and Victorian floods. Australian Journal of Emergency Management, The 27.1 (2012): 27.

[13] M. Dhiraj, A. Gross, and M. McGarry. Visual Social Media and Big Data. Interpreting Instagram Images Posted on Twitter. Digital Culture & Society 2, no. 2 (2016): 113-134.

[14] M. Clark-Ibez, Framing the social world with photo-elicitation interviews, American behavioral scientist 47.12 (2004): 1507-1527.

[15] M. David, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." (2011).

[16] F. Pedregosa, G. Varoquaux, A.Gramfort, V. Michel, B. Thirion, O.Grisel, M. Blondel et al. Scikit-learn: Machine learning in Python. Journal of machine learning research 12, no. Oct (2011): 2825- 2830.