Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises

Muhammad Imran
Qatar Computing Research Institute
mimran@qf.org.qa

Carlos Castillo
Qatar Computing Research Institute
chato@acm.org

Ji Lucas
Qatar Computing Research Institute
jlucas@qf.org.qa

Patrick Meier
Qatar Computing Research Institute
pmeier@qf.org.qa

Jakob Rogstadius
University of Madeira
jakob.rogstadius@gmail.com

ABSTRACT

An emerging paradigm for the processing of data streams involves human and machine computation working together, allowing human intelligence to process large-scale data. We apply this approach to the classification of crisis-related messages in microblog streams. We begin by describing the platform AIDR (Artificial Intelligence for Disaster Response), which collects human annotations over time to create and maintain automatic supervised classifiers for social media messages. Next, we study two significant challenges in its design: (1) identifying which elements must be labeled by humans, and (2) determining when to ask for such annotations to be done. The first challenge is selecting the items to be labeled by crowdsourcing workers to maximize the productivity of their work. The second challenge is to schedule the work in order to reliably maintain high classification accuracy over time. We provide and validate answers to these challenges by extensive experimentation on real-world datasets.

INTRODUCTION

Social media is increasingly used to communicate real-time information during crises. This user-generated content, which comprises text, imagery and video footage, can in some situations augment situational awareness. Analysis of messages posted on the Twitter microblogging service during disasters indicates that some messages refer to information relevant to disaster responders, e.g., infrastructure damage, needs and fatalities (Vieweg 2012). This analysis also shows that these messages are usually communicated more quickly than disaster information shared via traditional channels. For instance, the first tweet to report on the 2013 Westgate Mall attack[1] was posted within a minute of the initial onslaught.[2] However, social media streams also contain a significant amount of noise, and the majority of tweets posted during crises are not relevant let alone actionable. This was certainly true of the 35,000+ tweets posted in the immediate aftermath of the 2013 Pakistan Earthquake.[3]

Information overload during disasters can be as paralyzing to humanitarian response as the absence of information. The challenge is akin to the proverbial needle in the haystack problem; finding the needle—informative content—in real time within a rapidly growing stack of information. To identify valuable information, and to direct information to the right decision maker, messages need to be sorted into high-level, meaningful categories of information. Extensive research into classifiers for short text strings has mainly concluded that classification and information extraction from short texts on social media is significantly harder than for longer documents such as news articles and/or blog posts (Chenliang et al. 2012). In addition, in the disaster space, research has shown that while classifiers can be developed to work well during a single disaster, their performance drops significantly when reused in different but similar disasters (Imran et al. 2013d). This has made it difficult to build systems that incorporate pre-trained classifiers to be deployed during new events.

[1] We refer to the attack on September 2013 on the Westgate shopping mall in Nairobi, Kenya. http://iRevolution.net/2013/11/18/westgate-information-forensics

[2] http://www.ihub.co.ke/blog/2013/10/how-useful-is-a-tweet-a-review-of-the-first-tweets-of-the-westgate-attack

[3] We refer to the earthquake on 24 September 2013 with epicenter in the south of Pakistan. http://iRevolution.net/2013/09/27/results-of-micromappers-pakistan-quake

Real-time insights are important for emergency management, and the practitioner community has traditionally relied on manual approaches to filtering and classification of social media data. For instance, the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA) and other humanitarian organizations employ a Digital Humanitarian Network[4] of volunteers during humanitarian crises. They seek to gain additional situational awareness by having these digital volunteers monitor and filter social media, particularly Twitter. The same is true of the Digital Operations Center[5] at the American Red Cross, which activates over 150 certified digital volunteers to listen to social media during disasters.

Supporting such a workflow, the Ushahidi[6] system gained popularity in the disaster response community by enabling affected populations to send in reports, each of which is manually annotated by a crowd of human curators before it can be consumed by end users through a web front-end. Though the processing may be accurate, Ushahidi has serious issues with scalability since the inflow rate of information can reach thousands of messages per minute, easily surpassing the capacity of the human curators during mass disasters, even when hundreds of annotators are available (Morrow et al. 2011). This results in "filter failure", with an ever-growing backlog of unprocessed reports and with no way of knowing which reports to prioritize.

A semi-automatic system, CrisisTracker (Rogstadius et al. 2013), integrated manual processing of social media streams with automatic clustering. In their study, social media messages were clustered together by their textual similarity, and human curators were directed to annotate the larger clusters first. While this automated ranking and deduplication of reports greatly improved the *allocation of work* and reduced *redundancy*, it did not address the fundamental issue of increasing *work output*. In both Ushahidi and CrisisTracker, an average human annotator will process at most a few dozen items per hour (Rogstadius et al. 2013).

Even if a large enough workforce could somehow be amassed to keep up with the velocity of crisis information, having humans do something that can be automated is inefficient and a misallocation of scarce resources. This is especially true during prolonged crises, such as civil wars, or during the long recovery phase following major natural disasters. In these cases, dependency on human annotation adds a significant operating cost that cannot be sustained over time. At this processing rate, even with de-duplication strategies implemented as in CrisisTracker, hundreds of simultaneously active workers (thousands in total) would be needed to keep up with all social media activity during the most impactful disasters.

Problem: In order to ingest and process data in real-time, we adopt the Crowdsourced Stream Processing paradigm as described in (Imran et al. 2013a). Specifically, we use AIDR,[7] a system in which human assessors provide labels that are used to train an automatic classification system. In this setting, training data needs to be collected during an ongoing disaster, suggesting that as much labeling effort as possible should take place as early as possible. However, if there is significant temporal variance in the content stream, the labeling effort should be distributed more evenly over the duration of the crisis. In this work we aim to determine to what extent temporal effects are present in social media corpora collected during a single disaster, and what impact that these effects may have on classification accuracy. We then try to find an optimum for how crowdsourced annotation efforts should be allocated over the duration of a disaster.

Our contribution: In this paper we study an application in which a crowd of annotators provides training labels (i.e., classified messages) to train supervised classifiers that are applied to a message stream. Our hypotheses are that optimizing the annotation work requires: (i) new labels should be collected for each new crisis; (ii) no duplicate elements should be labeled; and (iii) active learning should be employed. We validate our hypotheses by experimenting on datasets from several crises. Moreover, we introduce different labeling strategies i.e., all the labels at the beginning, versus cumulatively adding labels as the crisis progresses, or creating a new set of labels for each stage of a crisis, and determine the best scheduling strategy for labeling. Overall, the paper solves key problems that lie in the interaction between crowdsourcing workers and the sub-systems that uses the labels to automatically learn to tag tweets, in particular with respect to selecting *which* tasks the crowdsourcing workers should do, and *when* should they do them.

RELATED WORK

Automatic classification of social media during crises. Previous work has proposed information management systems that incorporate automated content classifiers for use during disasters. These systems have relied on

[4] http://digitalhumanitarians.com/
[5] http://dell.to/yMHdGA
[6] http://www.ushahidi.com/
[7] http://aidr.qcri.org/

generic news classifiers applicable across a wide range of application domains (Abel et al. 2012), but not developed directly for short social media content. Other systems have relied on classifiers specifically developed for use during a particular disaster type (Yin et al. 2012). As these classifiers are static, they cannot be updated by end users if the classification needs or the underlying data change. Imran et al. (2013d) found that classifiers trained on data from one disaster suffered greatly reduced classification performance when applied to data of other disasters, even when the disasters were of the same type.

Supervised classification using AIDR. In (Imran et al. 2013b), the authors proposed AIDR (Artificial Intelligence for Disaster Response), a Crowdsourced Stream Processing (CSP) system which demonstrated how end users can be supported to develop automated classifiers based on their own information requirements during a specific event. AIDR thus removes the need to rely on pre-defined generic classifiers. Rather, AIDR is specifically designed to scale up the ability of crowds of human volunteers to classify large volumes of social media content during humanitarian disasters. In previous work, the authors showed that AIDR offers low latency, high throughput, high load adaptability, good cost effectiveness and good output quality (Imran et al. 2013a).

AIDR is designed to work with content from the Twitter microblogging service. It (1) collects crisis-related tweets using keywords or geographical boundaries; (2) it samples from those tweets a sub-set to be labeled; (3) it asks a group of crowdsourcing workers to provide labels for those; (4) it uses supervised learning train an automatic classifier based on the labels; and (5) it classifies the remaining data using that automatic classifier. Steps 2-5 are repeated periodically in order to capture more training data and to keep the classifier up-to-date.

From the user's perspective, AIDR works as follows. First, a user starts a new data collection through a web interface by filling in a set of disaster-related keywords to be tracked on Twitter, such as key location names, event-specific hashtags, prominent entities, etc. Regions can also be defined to collect geotagged content, and filters can be specified to limit the collection to messages in specified languages. The system also provides a dashboard for monitoring the collection status with metrics such as total processed items, time since collection started, and text and timestamp of the most recent message.

The user then specifies the label sets according to which the collected content should be classified. The user can either select from previously defined classification schemes or define new label sets, together with descriptive definitions of each mutually exclusive label within the set. Each label set corresponds to one classifier, and multiple label sets can be used simultaneously for one collection.

Next, a crowd of annotators provides training examples, by assigning labels to system-selected messages. This is done through a dynamic form that presents a message, the possible labels to apply, and their definitions. The training examples are then used by the system to train classifiers for incoming items. A subset of the training examples are held out for classifier evaluation, based on which the system reports classification accuracy scores back to the user. Social media messages that are collected by the system are classified using all active classifiers, using the best available model(s) at the time the message is received. Classification accuracy is thus affected not only by the size of the training data, by also by when that data is provided.

Temporal factors. (Mourão et al. 2008) analyzed temporal variance in text at the scale of years in two document corpora of scholarly articles. They found that classifiers trained on a time-local sample outperformed classifiers trained on the global corpus and greatly outperformed classifiers trained on a document sample from distant years. They showed the presence of three types of temporal effects: changes in class distribution, term distribution, and class similarity. These effects all reduce classification performance over time, by introducing new concepts, new features for existing concepts, or by moving decision boundaries between classes. In a follow-up work, (Rocha et al. 2008) introduced the concept of temporal context selection, where the goal is to determine the largest subset of available training data that minimizes the temporal effects.

Similar effects have been observed for much shorter time intervals in search query logs and in social media data collected during natural disasters. (Kulkarni et al. 2011) showed how both terms used in search queries and search hits relevant to specific search terms can vary greatly between days, indicating short-term presence of changes in class distribution and class similarity.

RESEARCH APPROACH AND FRAMEWORK

Social media activity during crises exhibits a variety of temporal patterns (Chowdhury et al. 2013). Similarly to the dynamic of Twitter hashtags as observed by (Romero et al. 2011), we consider two variables that describe these patterns: whether the crisis is expected or not, and for how long the crisis lasts.

The first variable entails the existence (or not) of a preparedness phase, in addition to response and recovery

phases. Predicted crises (e.g., hurricanes, floods due to heavy rainfall) allow members of the public and formal response agencies to prepare for impact. The second variable, the duration of the crisis, can be measured in days (most crises, including earthquakes and hurricanes), weeks (medium-term human conflicts including revolts), months or years (long-term human conflicts, pollution disasters such as oil spills, etc.). We remark that the sense of duration of a crisis is open to interpretation: it is hard to define when they end.

For purposes of the present research, we make an attempt to stay as general as possible, but given that several of our findings are related to the time-sensitive nature of these data, we focus on a specific (but broad) family of crises: disasters created by atmospheric phenomena that are *announced* and *span several days*. Details on our datasets are given on Table 1.

Dataset	**Description**	**Size**
Joplin 2011	The dataset was collected using the hashtag (#joplin) during the Joplin tornado that struck Joplin, Missouri, USA on Sunday, May 22, 2011.	206,764
Sandy 2012	The dataset was collected using keywords (hurricane sandy, frankenstorm, #sandy) during the Sandy hurricane that hit Northeastern US on October 29, 2012.	4,906,521
Oklahoma 2013	The dataset was collected using 38 keywords (including: oklahoma tornado, oklahoma storm, oklahoma relief, oklahoma volunteer) during the Moore tornado that struck Moore, Oklahoma, USA on the afternoon of May 20, 2013.	2,742,588

Table 1: Datasets used in our research. The size corresponds to the total number of tweets on each dataset.

The profile of volume of messages (tweets) with phases (S1-S4) vs. time is shown in Figure 1. In each dataset, we have identified a series of phases based on the volume of tweets. In our experiments, for consistency we do not attempt to divide our time in days. Instead, we identify different phases of a crisis corresponding to entire days and having comparable numbers of tweets on each one. Table 2 shows the phases we identify in each of our datasets.

The "*pre*" phase corresponds to the preparedness phase, before the main effects of the natural hazard are experienced. The "*impact*" phase corresponds to the period in which the main effects are felt, and is accompanied by the larger volume of tweets. The "*post*" phase comes last and corresponds to the response and recovery phase. In case of Joplin dataset, in addition to the *pre*, *impact* and *post* phases, we identified *pre'*, *impact'*, and *post'* phases. These phases show a significant amount of messages that we identified after the *post* phase of the Joplin dataset. In general, disasters like tornados and hurricanes exhibit different pre-disaster profile, as it can also be observed in our datasets. For this reason, to have a comparable number of messages across all the datasets for phase S1, we combined *pre* and *impact* phases of Joplin and Oklahoma datasets.

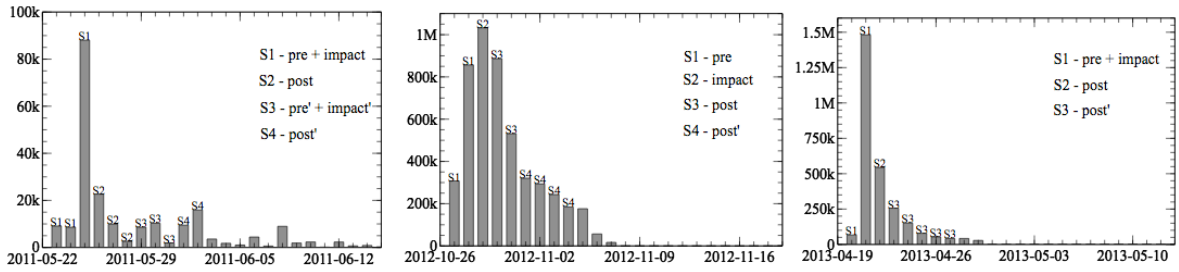

Figure 1. Number of tweets per day in our datasets: Joplin (left), Sandy (center), and Oklahoma (right). The scale of the time axis is same in the three plots.

Classification Tasks: Informative vs. Not-Informative, Donation vs. Not-Donation

To determine the temporal effects and their impact on classification accuracy, we consider two concrete classifiers in the crisis-response domain. The first classifier: "informative vs. not informative", helps detect messages that contain situational awareness information, for example, infrastructure damage, casualities, donation requests or offers (the "informative" class) vs. those that do not (the "not informative" class).

The question asked to human assessors follows (Imran et al. 2013c), in the sense of using three coding options: (1) Informative, if a message contains disaster-related information and can be used for situational awareness;

	Joplin			Sandy			Oklahoma		
Phase	Volume	% inf.	#days	Volume	% inf.	#days	Volume	% inf.	#days
S1	105,760	69%	3	1,164,269	42%	2	1,548,031	56%	2
S2	35,145	58%	3	1,032,575	39%	1	543,102	58%	1
S3	20,790	65%	3	1,417,606	54%	2	580,987	58%	5
S4	25,426	57%	2	1,042,549	49%	4	N/A	N/A	N/A

Table 2. Crisis phases, indicating the number of tweets on each one, and the number of days of the fraction of informative messages (according to crowdsourcing workers) on each phase.

(2) Personal, if a message does not convey useful information to others and is only of interest to the author's family/friends; and (3) Other, if a message is not in English or it cannot be classified. A consensus of three workers was required for a tweet to finalize its label, and we merged the responses (2) and (3) into a single "not informative" class.

Examples of informative messages in our data as identified by crowdsourcing workers are:

- *"@NYGovCuomo orders closing of NYC bridges. Only Staten Island bridges unaffected at this time. Bridges must close by 7pm. #Sandy #NYC",*
- *"Tornado Warning issued May 26 at 4:59PM CDT expiring May 26 at 5:30PM CDT by NWS New Orleans"*

Class distribution changes (Mourão et al. 2008) are evident in the data. As we can see in Table 2, in the case of the Oklahoma dataset there is consistency in the fraction of informative messages over time, in the Joplin and Sandy datasets we observe significant variations of up to 12 to 15 percentage points from one stage to another.

We built a second classifier, which is more fine-grained and operates over the informative messages. It asks if a message is related to emergency response resources that can be donated, including goods such as food, water, medicines, etc. as well as services such as volunteer work. This classifier outputs two classes, which we call "donation" and "not donation". Given the fact that the prevalence of the donation-related messages in the data is smaller than the prevalence of informative messages as the former is a sub-set of the latter, it is interesting to see the temporal effects variation in the case of donation classifier as compared to the informative classifier.

Examples of the donation messages in our data, as identified by crowdsourcing workers, are:

- *"400 Volunteers are needed for areas that #Sandy destroyed",*
- *"Help is on its way to #Joplin. 24 food truck with donations".*

In both classifiers, to collect training data we used the CrowdFlower[8] platform and required three workers' agreement on a label.

Learning scheme: random forests with feature selection

In previous off-line tests of classification of messages during natural disasters (Imran et al. 2013c) we performed extensive experimentation with different feature extraction and selection strategies, as well as different learning schemes. Features that consistently yielded good results were word unigrams and bigrams (single words or sequences of two consecutive words). We performed feature selection using the information gain criterion as implemented in WEKA 3.7.6,[9] keeping the top 500 features.

As a learning scheme, we used random forests (Breiman, 2001), which are ensembles of decision trees using bagging and random feature selection. We used the implementation in WEKA 3.7.6 with default parameters, which are number of trees in the forest (default = 10), S = 1 (i.e., seed for random number generator), and no maximum depth of the trees. We remark that our results are to a large extent independent of the learning method used, as long as the learning method is strong enough to separate the classes of interest.

[8] http://crowdflower.com/

[9] http://www.cs.waikato.ac.nz/ml/weka/

We measure the classification quality using AUC[10] (Area under the Receiver Operating Characteristic curve), which is a standard measure of classification performance. It can be interpreted as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. In this context, the score of an item is the probability of belonging to the positive class as determined by the automatic classification system. The AUC is a value between 0.0 and 1.0, where 0.5 meaning random classification and 1.0 meaning perfect classification.

LABELING TASK SELECTION

Labeling consumes a limited resource: the time of crowdsourcing workers. Even when done on a voluntary basis, the time of crowdsourcing workers is neither unlimited nor free of costs. We thus assume there is a certain budget in terms of a maximum number of items that can be manually labeled during a crisis. We then study how to select the elements that will be labeled in order to maximize the accuracy of the classifier built using the labels. In this section we consider three central aspects of this selection process. First, it is a necessity to have fresh labels for every crisis. Second, we should avoid labeling items that closely resemble those we have already labeled, both because it wastes the assessor's time but also because it may be misleading during the evaluation. Third, we should prioritize items that are likely to lead to larger improvements in the classifier, a technique known as *active learning*.

Crisis-specific labels are necessary

Reuse of labels from different but similar crisis situations could significantly reduce the labeling effort and cost. To examine the performance in such *transfer* scenarios, training a classifier on two crises and testing the third one, we performed 3 experiments using the Joplin, Sandy and Oklahoma datasets. In this case, we used the entire labeled data available for each crisis (as will be explained later, this consists of 5,000 labeled tweets for Joplin, 5,000 for Sandy and 4,000 for Oklahoma) classified by human assessors as informative or not-informative, which is the main task we study.

The performance (AUC) we obtain is 0.52 (train on Joplin, test on Sandy), 0.56 (train on Joplin, test on Oklahoma) and 0.53 (train on Sandy, test on Oklahoma). We remark that a random classifier would have obtained an AUC of 0.50. Many reasons that can significantly damage the performance include, for example, the use of contrasting vocabulary even during the same type of crisis or the fact that there are differences in the public concerns, affected infrastructure, etc.

Data redundancy leads to overestimates of classification quality

In AIDR, we employ aggressive deduplication strategies. Duplication in social media content is highly prevalent and an analysis of five Twitter datasets in (Rogstadius et al. 2011) found that 29-47% of tweets were retweets, and 60-75% of tweets are duplicates or near duplicates of another tweet, as determined by a clustering algorithm. Near duplicates occur for instance when different users post tweets with semantically equivalent content after reading the same news article, or when doing a minor change before reposting a message rather than retweeting.

The presence of duplicate items in a training set can cause classifier bias, and for a fixed number of training samples it will reduce the number of concepts the learner is exposed to. Deduplication is not only done to maximize gains in accuracy but also to improve the user experience of workers. Presenting the same examples over and over to crowdsourcing workers, or closely similar ones, is likely to alienate them or give them the impression that the system is not responding to their training.

In AIDR, the deduplication process is performed using an online algorithm. First, the tweets are preprocessed by removing all "RT @username" prefixes and replacing all URLs by the token "_URL_". Second, tweets are converted to unigrams and compared against the latest 50 tweets seen (a time-sorted bounded buffer is used); if the similarity[11] with any of them is larger than 0.5, the tweet is declared to be a duplicate.

To examine the effects on learning performance of performing deduplication, we crowdsourced a random sample of 5,000 tweets from the Sandy dataset (2,000 from phase S1, 1,000 from each of the S2-S4 phases as

[10] See e.g. https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve

[11] We measured the Jaccard similarity of the two sets, $J(A,B) = |A \cap B|/|A \cup B|$

shown in Table 2) without applying deduplication, for the task of informative vs. not informative. To get labels for deduplicated data, we again crowdsourced same number of tweets taken from each phase after applying deduplication and presenting the same set of categories for labeling.

Results are shown in Table 3, where we can see that all cases show higher AUC values when not performing de-duplication. This is due to the reason that both training and test datasets contain duplicate messages. Having common elements in the training and testing sets causes an artificial increase in the reported effectiveness of the system. Ideally, training and testing data should not overlap.

Phase	**Train**	**Phase**	**Test**	**AUC (without de-duplication)**	**AUC (with de-duplication)**
S1 (pre)	1500	S1 (pre)	500	0.78	0.74
S1 (pre)	500	S1 (pre)	500	0.73	0.72
S2 (impact)	500	S2 (impact)	500	0.80	0.72
S3 (post)	500	S3 (post)	500	0.79	0.73
S4 (post')	500	S4 (post')	500	0.70	0.64

Table 3. Results with and without de-duplication on Sandy dataset. Without de-duplication, results tend to overestimate the quality of the classifier.

Passive vs. active learning

In supervised learning, usually a learning algorithm receives an input set of labeled data which is outside the algorithm's control; this is called *passive* learning. In *active learning*, the algorithm itself decides which unlabeled items should be labeled to be used as future training examples. At a high level, an active learning method tends to pick items for which it is unsure about their classification examples that are close to the decision boundary and for which the labels are maximally informative. We test the effects of the type of learning on the output quality as labels are received. Ideally, we would like to attain a high level of quality (e.g. high AUC) with as few labels as possible.

For these experiments, we obtained 2,000 crowdsourcing labels for phase S1, and 1,000 labels for each of the following phases (S2-S4) using randomly selected deduplicated tweets from each crisis. Next, we simulated the performance of the system under passive and active learning. We injected a stream of tweets into the system in chronological order, after which the tweets were labeled using a process that simulated the arrival of the labels. To avoid testing the system's performance on a very small test set, first we inject all the labels allowing the system to select every fifth labeled tweet as part of an evaluation set. Second, the remaining labels were injected again executing the learning process at every twentieth training example. Figure 2 shows the results of our experiments for all the crises.

Results show that the active learning approach dominates and AUC tends to stabilize earlier (i.e., with fewer training examples) than the passive learning approach. We note that in the case of Joplin the AUC reaches an acceptable level with fewer training examples than Sandy and Oklahoma, possibly a consequence of the higher prevalence of the positive (informative) class. Overall, the system reaches an acceptable AUC (i.e., > 0.75) after approximately 600 labels. In the context of online machine learning approach, which we also follow in AIDR, active learning technique is more suitable, because more labeled items could significantly increase the overall performance. Though, the requirement to include more labeled examples can depend on many other factors such as the complexity of categories, class distribution, or term distribution, etc.

LABELING TASK SCHEDULING

Once we have decided the strategy to select *which* messages to label, the next question is *when* to label them. Assuming a fixed number of labels, we could for instance use this labeling budget completely at the beginning of the crisis. However, it might be possible that the characteristics of tweets change over time and thus the classifier needs to be updated with new labels. In that sense, a gradual labeling strategy may be advisable. In this section we evaluate these alternatives experimentally.

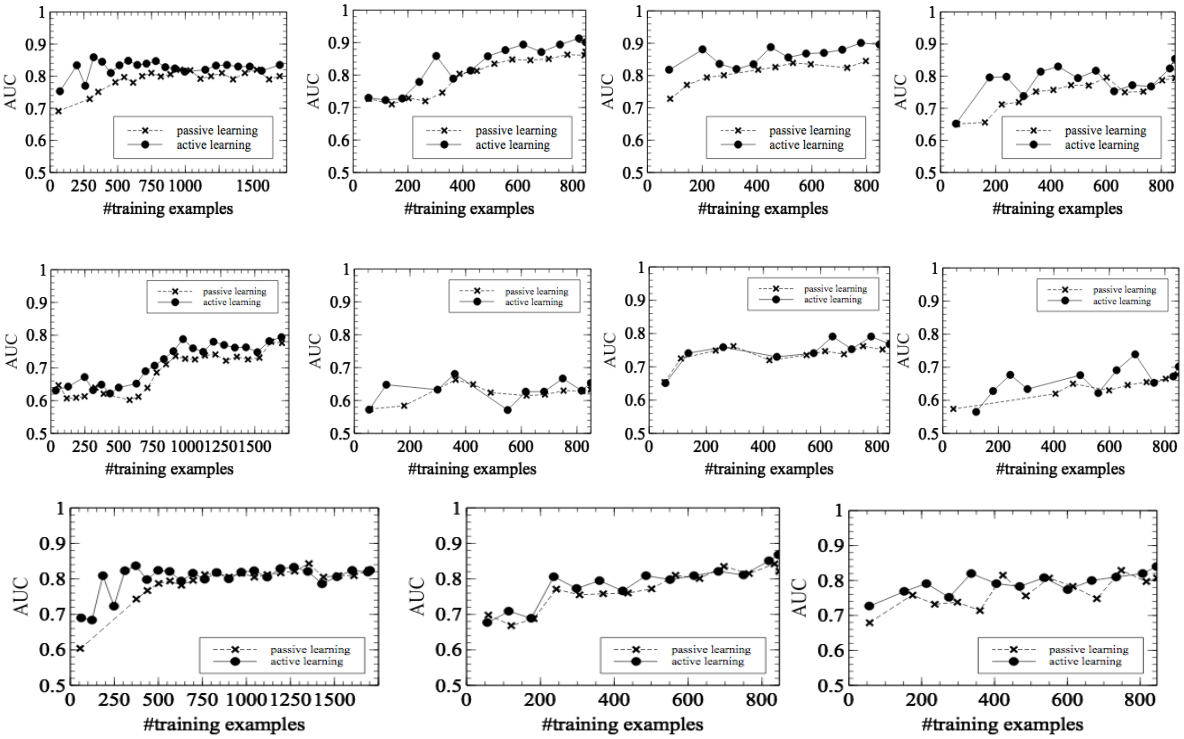

Figure 2. Results of output quality using passive and active learning, measured using AUC, vs. number of training labels. Each column corresponds to one phase starting from S1 (from left) to S4 (In the Oklahoma dataset there are only 3 phases). Datasets are: Joplin (top), Sandy (middle) and Oklahoma (bottom).

Scheduling strategies

For concreteness, let us consider the following three scenarios for a crisis spanning three periods of time (S1, S2, and S3). We note that the three scenarios below have the same cost in terms of the number of crowdsourcing labels necessary. Which one should yield better quality over time?

- *All-at-once*: Obtain 1,500 labels on S1 and use all of them for training.
- *Cumulative*: Obtain 500 labels in each of S1, S2, and S3, and use all the labels available up to each period for training (500 labels on S1, 1,000 on S2, and 1,500 on S3).
- *Independent*: Obtain 500 labels in each of S1, S2, and S3 and use only the most recent labels for training on each period, discarding old labels (500 labels on each of S1, S2, and S3).

Experimental results

In each crisis, we compared *all-at-once* vs. *cumulative* vs. *independent* labeling. We used the same labeled data that was obtained during our first round of experiments. Results are shown in Figure 3 for Joplin, Sandy, and Oklahoma respectively. From the results, we do not observe any aging effects. The *all-at-once* approach dominates over cumulative and independent approaches, as more data and thus better classifiers become available earlier. When repeating the experiments with half of the training data (*informative-50%*), we see a decrease in AUC across all methods, but the all-at-once approach continues to dominate.

Next, we look at a different labeling task, which is the one of labeling donation-related tweets. In this case, aging effects are evident. In the case of Joplin and Oklahoma, the *cumulative* strategy is better than the *all-at-once* because the latter rapidly decreases in accuracy. In the case of Sandy, the *all-at-once* strategy is not even a possibility, because less than 5% of donation-related tweets are found in the first stages of the crisis: too few to create a reliable classifier.

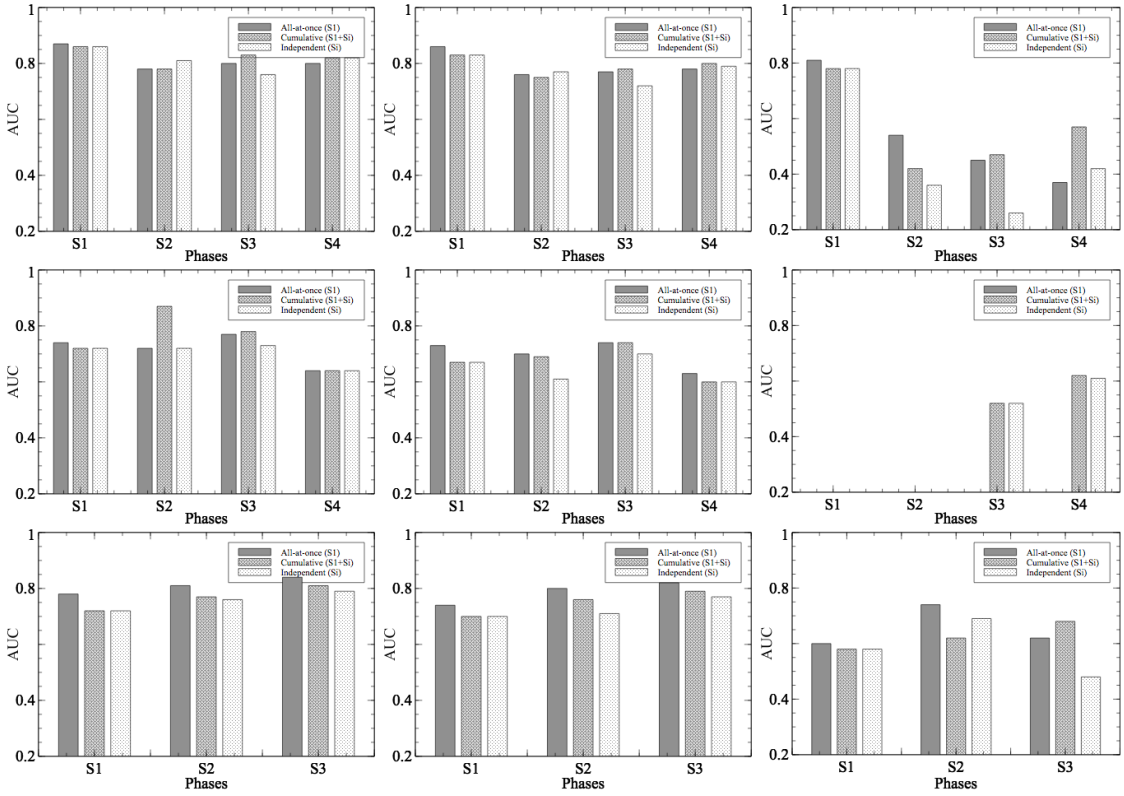

Figure 3. Results of output quality, measured using AUC for the task of identifying informative messages with all (left) and 50% (center) of the training data, and the task of identifying donation-related messages (right). Datasets are: Joplin (top), Sandy (middle) and Oklahoma (bottom).

DISCUSSION

Making sense of messages posted in social media requires the interaction of human intelligence and automatic processing. Currently, neither can do this job by itself. By a large margin, it is infeasible for even a large group of volunteers to keep up with the volume of tweets posted during a crisis in Twitter, especially during the *impact* phase. Automatic processing, in our case automatic classification, may yield very poor classification accuracy when labels are not current. A solution that integrates both aspects is promising in principle. However, there are aspects of such a solution that need to be designed with care, particularly in the areas where the two types of elements in the system (human and machine) interact. In this paper we have focused on how to better use the time of the crowdsourcing workers or volunteers, and report several findings in this regard.

Removing duplicates is extremely important because there is much redundancy. This redundancy means that messages in the testing portion of the labeled data may be too similar to messages in the training part, leading to overestimates in the classification performance. Our experimental results clearly show the artificial increase in performance due to the overlapping messages in the training and testing sets.

Label ageing/decay may or may not be a problem in the timeframe we studied (in the order of days), depending on the specific task. There are class distribution changes and term distribution changes. The latter are evident when detecting donation-related messages, which appear later during a crisis, but do not seem to strongly affect the more generic "informative" vs. "not informative" classification. However, in the timeframe we studied, the all-at-once approach performed better than the others.

Active learning (as expected) yields better classification accuracy with fewer labels, but not by a large margin. It is interesting to observe that classification accuracy varies as time passes. That is, as an actual crisis unfolds, and with the availability of more labels, the learning process should increase overall performance.

FUTURE WORK

Changes in the distribution of categories and the distribution of terms within each category depend on the specific categorization being used and on the period of time under consideration. Our work can be extended to

other categorizations (beyond the ones we used in this paper) and into longer periods of time.

Additionally, the measure of quality we use may not be what end users experience. For instance, if clustering is applied on the output of AIDR (as in the case of CrisisTracker), the accuracy at the cluster level is more important than the accuracy at the message level. Bias could be introduced if messages that are often repeated were somehow easier or harder to classify automatically.

Finally, the concerns of the designer of a system such as the one we have described do not relate exclusively to output quality. They also involve human aspects, especially if the crowd is composed of volunteers. All other things equal, volunteers may find repetitive or difficult tasks off-putting, which may lead them to abandon the platform in the long term. Ideally, our metrics should include measures of user engagement that help identify possible shortcoming in terms of keeping a community of volunteers coming back to help in future crises.

REFERENCES

1. Abel, F., Hauff, C., Houben, G-J., Stronkman, R. and Ke Tao. (2012) Semantics + filtering + search = twitcident. Exploring information in social web streams. *HT '12*, New York, NY.
2. Breiman, L. (2001) Random forests. *Machine learning*, 45, 1, 5-32.
3. Chenliang, L., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A. and Lee, B-S. (2012) TwiNER: named entity recognition in targeted twitter stream. *SIGIR '12*, Portland, OR.
4. Chowdhury, S. R., Imran, M., Asghar, M. R., Amer-Yahia, S. and Castillo, C. (2013) Tweet4act: Using Incident-Specific Profiles for Classifying Crisis-Related Messages. *ISCRAM '13*, Baden-Baden, Germany.
5. Imran, M, Ioanna L. and Castillo, C. (2013) Engineering Crowdsourced Stream Processing Systems. *arXiv preprint,* arXiv:1310.5463 (a).
6. Imran, M., Castillo, C., Lucas, J., Meier, P. and Vieweg, S. (2014) AIDR: Artificial Intelligence for Disaster Response. *WWW '14*, Seoul, Korea (b).
7. Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F. and Meier, P. (2013) Extracting information nuggets from disaster-related messages in social media. *ISCRAM '13*, Baden-Baden, Germany (c).
8. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F. and Meier, P. (2013) Practical Extraction of Disaster-relevant Information from Social Media. *WWW '13 Companion*, Rio de Janeiro, Brazil (d).
9. Kulkarni, A., Teevan, J., Svore, K. M. and Dumais, S. T. (2011) Understanding temporal query dynamics. *WSDM '11*, New York, NY.
10. Morrow, N., Mock, N., Papendieck, A. and Kocmich, N. (2011) *Independent Evaluation of the Ushahidi Haiti project*, Development Information Systems International.
11. Mourão, F., Rocha, L., Araújo, R., Couto, T., Gonçalves, M. and Meira, W. Jr. (2008) Understanding temporal aspects in document classification. *WSDM '08*, New York, NY.
12. Rocha, L., Mourão, F., Pereira, A., Gonçalves, M. A. and Meira, W. Jr. (2008) Exploiting temporal contexts in text classification. *CIKM '08*, Napa Valley, CA.
13. Rogstadius, J., Kostakos, V., Laredo, J. and Vukovic, M. (2011) A real-time social media aggregation tool: Reflections from five large-scale events. *CSCWSmart Workshop at ECSCW '11*, Aarhus, Denmark.
14. Rogstadius, J., Teixeira, C., Vukovic, M., Kostakos, V., Karapanos, E. and Laredo, J. (2013) CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness. *IBM Journal of Research and Development*, 57, 5, 4:1-4:13.
15. Romero, D. M., Meeder, B. and Kleinberg, J. (2011) Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. *WWW '11*, Hyderabad, India.
16. Turaga, D., Andrade, H., Gedik, B., Venkatramani, C., Verscheure, O., Harris, J. D. and Jones, P. (2010) Design principles for developing stream processing applications. *Software – Practice & Experience*, 40, 12, 1073-1104.
17. Vieweg, S. (2012) *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications*. Doctoral dissertation, University of Colorado at Boulder.
18. Yin, J., Lampert, A., Cameron, M., Robinson, B. and Power, R. (2012) Using Social Media to Enhance Emergency Situation Awareness. *Intelligent Systems, IEEE*, 27, 6, 52-59.