Enabling Rapid Classification of Social Media Communications During Crises

Muhammad Imran
Qatar Computing Research Institute, HBKU
Doha, Qatar
mimran@hbku.edu.qa

Prasenjit Mitra
The Pennsylvania State University,
University Park, PA, USA
pmitra@ist.psu.edu

Jaideep Srivastava
Qatar Computing Research Institute, HBKU
Doha, Qatar
jsrivastava@hbku.edu.qa

ABSTRACT

The use of social media platforms such as Twitter by affected people during crises is considered a vital source of information for crisis response. However, rapid crisis response requires real-time analysis of online information. When a disaster happens, among other data processing techniques, supervised machine learning can help classify online information in real-time. However, scarcity of labeled data causes poor performance in machine training. Often labeled data from past event is available. Can past labeled data be reused to train classifiers? We study the usefulness of labeled data of past events. We observe the performance of our classifiers trained using different combinations of training sets obtained from past disasters. Moreover, we propose two approaches (target labeling and active learning) to boost classification performance of a learning scheme. We perform extensive experimentation on real crisis datasets and show the utility of past-labeled data to train machine learning classifiers to process sudden-onset crisis-related data in real-time.

Keywords: Social media, tweets classification, domain adaptation, disaster response

1. INTRODUCTION

In the last few years, the use of social media platforms during disasters and emergencies has increased. In particular, microblogging platforms such as Twitter provide active communication channels during the onset of mass convergence events such as natural disasters (Palen et al., 2009; Hughes et al., 2009; Starbird et al., 2010; Vieweg et al., 2010). Studies show that Twitter has been used to spread news about casualties and damage, donation offers and requests, and alerts, including multimedia information such as videos and photos during crises (Cameron et al., 2012; Imran et al., 2013a; Qu et al., 2011). Many studies show the significance of this online information (Vieweg et al., 2014; Sakaki et al., 2010; Neubig et al., 2011) for crisis response and management. Moreover, it has been observed that these messages are usually communicated more quickly than disaster information shared via traditional channels such as news websites, etc. For instance, the first tweet to report on the 2013 Westgate Mall attack was posted within a minute of the initial onslaught.[1] Given the importance of crisis-related messages for time-critical situational awareness, disaster-affected communities and professional responders may benefit from using an automatic system to extract relevant information from social media.

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

Among other benefits that encourage responding organizations to use social media data is the timeliness of information when there are no other information sources available, especially in the beginning of a crisis situation (Tapia et al. 2013). For this reason, to enable rapid crisis response, real-time insights of an ongoing situation play an important role for emergency responders. To identify informational, actionable, and tactical informative pieces from a growing stack of social media information and to inform decision-making processes as early as possible, messages need to be processed as soon as they arrive. Given the large volume of messages, we need to classify them. That is, we need to put them in different informational categories such as food needs, supplies requests; financial support requests, logistics, etc. so that disaster-response professionals can quickly examine each bin to identity urgent needs.

Different approaches can be employed to filter and classify these online messages. For instance, many humanitarian organizations use the Digital Humanitarian Network (DHN)[2] of volunteers to analyze messages one by one to find useful information for disaster response. However, given the amount of information that needs to be processed, and the scarcity of volunteers, we would ideally like the messages to be categorized automatically, and volunteers to use their time to perform higher-order tasks. Despite advances in natural language processing, full automation is still not feasible.

In this paper we propose to use a hybrid approach in which both humans and machines work together to perform complex tasks (e.g. classification of tweets). Among other automatic processing techniques, most automatic classifiers that achieve high accuracy in solving different classification tasks are based on supervised machine learning techniques where humans provide a set of training samples consisting of positive and negative examples for each classification category. For instance, a semi-automated system having similar characteristics to DHN is AIDR (Artificial Intelligence for Disaster Response) (Imran et al., 2014).

The AIDR platform collects event-specific data (using user-defined queries) from the Twitter streaming API, and uses supervised machine learning techniques to classify messages into user-defined categories or bins. AIDR is trained by humans which then automatically process and classify messages at high-speed using a supervised classification technique. AIDR or any other similar system that performs automatic classification requires human-labeled example examples pertaining to each category. Scarcity of such human-labeled data results in poorer classification models or it delays the machine training process. Gathering human-labeled data for training classifiers is a hard problem because human annotators find it a boring and laborious task, especially if they are doing it in large numbers. Moreover, the task becomes more challenging under time-critical situations where the need to make-sense of large data is in high-demand. However, what if we use human-labeled data from past events. For example, AIDR has been used to collect data from similar events in the past and has annotated data that can be used, if they are found useful. If we can reuse the existing annotations from AIDR, then we can also improve the accuracy significantly resulting in a much better model.

In this work, we utilize labels from past crises to train machines so that they can classify messages from new crises. However, the problem is when multiple such past crises exist, we need to choose which ones are useful and which are not. The traditional machine learning premise is that we should use as much relevant training data as we have. However, should we use labeled messages from different languages (e.g. data from two earthquakes events in

[2] http://digitalhumanitarians.com/

different languages) from the same event type to train? Are all the datasets from the same event, for example an earthquake, relevant for the next earthquake? Because the different datasets originate from different parts of the world, they use different languages or mix of languages, etc. Datasets for a similar event, e.g., earthquake may not be useful from one event to another. We wanted to examine the datasets to see if existing datasets and their tagged examples helps.

Provided there exists crisis-related collections along with human-tagged data (e.g. in case of AIDR), we specifically examine the following questions empirically. 1) Can we use the past data to build machine learning models? 2) Will using the past data improve the models? 3) Should we use all the labeled data?

To get answers to the above questions, we train classifiers using different combinations of existing data and examine on unseen test data how they perform in order to address these questions. We show that in most cases data from the same domain are very useful. A few exceptions exist. For example, Italian tweets improved the classification performance of tweets from Spanish-speaking countries but not English-speaking countries[3]. However, beyond language, there could potentially be other variations due to which we need to be careful in choosing existing data for training. For example, we believe that variations in dialects, vernaculars, season, geography, urban/rural divide, development status of countries etc. could potentially render the datasets and the discussions from the same type of event to be different.

To the best of our knowledge, our work is the first to use existing tagged information in conjunction with information tagged for the specific event to train classifiers at this scale, and show that using the old data helps improve performance in most cases. Because our evaluation found a few anomalous cases, we recommend that before deployment, we need to validate the impact of the additional datasets on the performance of the classifier using a small test set before including the training data to create the classification model. To achieve maximum performance, we should not add the training datasets that cause the classifiers to lose accuracy during this validation step.

Having established that the labeled data from same domains is generally useful, next we ask the following question. Can we use data from one domain, e.g., earthquake, to train models for another domain, e.g., floods? In computer science, this is a well-known problem of domain adaptation (Daume et al., 2006). In supervised classification setting, one of the basic assumptions in learning new classifiers is that the training and test sets instances are drawn from the same data distribution. If the training and test sets differ substantially then it causes problems for the learning scheme to generalize. However, to deal with the inherited problem of labels scarcity, we aim to investigate how useful the labels of past crisis events can be for classifying a target crisis.

Moreover, once we learn a best setting for a given event type (e.g. best performing models using past data), we aim to improve the classification accuracies of our baseline models using different approaches. Specifically, in this paper, we try two approaches as follows. First, based on our findings that event-specific labeled data always help achieve better performance, we aim to predict labels for target events unlabeled data and then include the items with high classification accuracy in the training set of the model. The second approach we use is to employ the active learning technique to select items from a target event for which the learner's confidence is low and then ask human annotators for their labels. This approach helps classifiers learn complex

[3] This observation is a fascinating example of big data science. This result seems to point out that certain languages are closer to one language than others. And, that there is value in cross-language training of classification models.

classes using fewer numbers of labels (i.e. cost reduction). Results from both experiments show improvements in classifiers performance.

The rest of the paper is organized as follows. In the next section, we describe real-time classification approach which is an application area of our current work. Datasets details and then experimental setup sections provide details regarding what datasets we use and our experimental plan. We discuss results in the discussion section and elaborate related studies in the related work section. Finally, the paper is concluded in the conclusions section.

2. REAL-TIME CLASSIFICATION APPROACH

To be useful and actionable for emergency managers, information must be delivered to them in a timely fashion during a crisis situation (Tapia et al. 2013). In the case of social media data, this timeliness is achieved by using a real-time stream-processing paradigm (e.g. Imran et al., 2013b), in which data items are processed as soon as they arrive. Stream processing is different from batch processing, in which an archive with the information to be analyzed preexists and the processing is performed in a retrospective way.

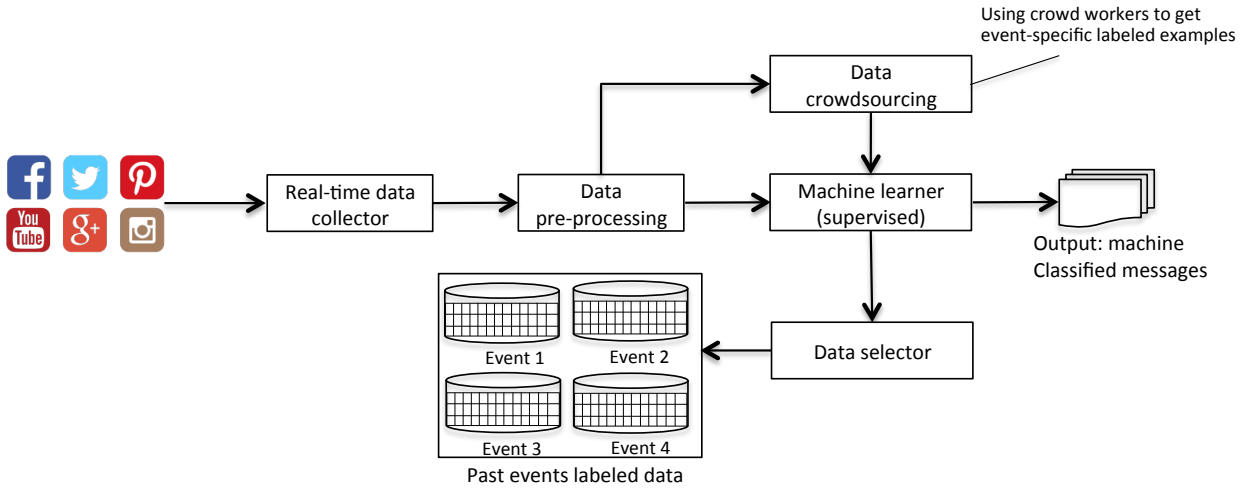

Figure 1: Pipeline for the classification of messages in real-time using supervised machine learning setting in conjunction with past events data

Different data processing techniques can be used for real-time analysis of data streams (Imran et al., 2013b). In this paper we use supervised classification techniques. Figure 1 shows the proposed approach, which is an extension of the AIDR approach, in the supervised machine learning setting. First of all, the data is collected from a number of social media sources using the real-time data collector component, followed by data pre-processing component, which performs required pre-processing steps for training machine learning classifiers. In this stream processing setting, we have two core components called data crowsourcing and machine learner. Data crowdsourcing responsible to get fresh labeled examples from human, whereas, the machine learner component uses those labeled examples to train machine learning classifiers. Specifically, humans train machines by providing event-specific labeled examples. However, human labeling cannot scale to the data volumes typical of large-scale crises, and is usually done on a sample of the input data. Whereas, automatic labeling by machines can overcome this issue, for example, by using human labeled data to train a supervised classification system. In this hybrid approach event-specific training data provided by humans is used to train and re-train an

automatic classification system (e.g. Imran et al., 2014). Availability of the human labeled messages is a core aspect in this processing pipeline. However, as described earlier during the sudden onset of a crisis situation, especially in the early hours when no other means of information exist, scarcity of human-labeled data introduces a high latency to process and produce useful results for crisis responders.

To overcome this bottleneck, next we study the usefulness of past-labeled data available from previous crises. We perform extensive experimentations on a number of real crisis datasets (described next) and learn how labeled data from past crisis events can be utilized to process a new target crisis.

3. PROBLEM FORMULATION AND CRISIS DATASETS

In this section, we first formally define our problem and classification setting and then describe an experimental framework.

3.1 Problem formulation

We consider this as multi-class classification problem in different domain adaptation settings, which we formally define as follows:

Given a domain D, which consists of two parts $\big(X, P(X)\big)$, where X represents the feature space and $P(X)$ represents marginal probability distribution. Now, given a training data set (X_i, Y_i), where $x_i \in X$ is the i^{th} feature vector and $y_i \in \{1, \dots, K\}$ class labels, in the multi-class classification problem, the aim is to learn a predictive function $f(.)$ using the feature vectors and labeled pairs $\{X_i, Y_i\}$.

From the above definition, we represent D_s as the ***source domain*** data where $D_s = \{(x_{s1}, y_{s1}), \dots, (x_{sn}, y_{sn})\}$ where $x_{si} \in X_s$ is the i^{th} instance of D_s and $y_{si} \in Y_s$ is the corresponding class label for x_{si}. Similarly, we represent D_T as the ***target domain*** data where $D_T = \{(x_{T1}, y_{T1}), \dots, (x_{Tn}, y_{Tn})\}$ where $x_{Ti} \in X_T$ is the i^{th} instance of D_T and $y_{Ti} \in Y_T$ is the corresponding class label for x_{Ti}. Now given multiple source domains D_s and a target domain D_T, in this work we aim to learn a predictive function $f_T(.)$ for the target domain data classification by using the information from both D_s and D_T source and target domains respectively.

In this work, we always consider $D_s \neq D_T$, but $T_s = T_T$ (i.e. source and target domains tasks are same). When both source and target events belong to the same crisis type (e.g. both D_S and D_T are earthquake events), we represent them as an ***in-domain*** case. And, we use ***cross-domain*** to represent when both source and target events have different crisis types (e.g., D_s comes from an earthquake event and D_T comes from a flood event).

3.2 Datasets

We use a combination of data collected by the AIDR platform and from the CrisisLexT26 dataset (Olteanu et al. 2015). Both datasets correspond to social media messages from Twitter posted during different crises that took place in 2012, 2013, and 2015. We selected 11 crises of two types: earthquake (5 crises) and floods (6 crises). Table 1 lists the crises along with other salient details. AIDR uses volunteers during the onset of a crisis situation to label crisis-related

messages. However, CrisisLex used paid crowdsourcing platforms for human labeling. In the datasets, each crisis corresponds to 800+ tweets annotated using the "Information Type" annotation scheme, which classifies tweets into the following categories (each tweet assigned only one of the following categories):

- **Affected individuals:** deaths, injuries, missing, found, or displaced people, and/or personal updates.
- **Infrastructure and utilities:** buildings, roads, utilities/services that are damaged, interrupted, restored or operational.
- **Donations and volunteering:** needs, requests, or offers of money, blood, shelter, supplies, and/or services by volunteers or professionals.
- **Caution and advice:** warnings issued or lifted, guidance and tips.
- **Sympathy and emotional support:** thoughts, prayers, gratitude, sadness, etc.
- **Other useful information:** not covered by any of the above categories.
-

Crisis name (short name)	Language	Date happened	Crisis type	# of labels
Italy earthquake (ITEQ)	IT + ENG	20-May-2012	Earthquake	911
Costa Rica earthquake (CREQ)	ES + ENG	05-Sep-2012	Earthquake	866
Guatemala earthquake (GUEQ)	ES + ENG	07-Nov-2012	Earthquake	905
Bohol earthquake (BOEQ)	ENG	12-Oct-2013	Earthquake	943
Nepal earthquake (NEEQ)	ENG	25-Apr-2015	Earthquake	2,812
Philippines floods (PHFL)	ENG	01-Aug-2012	Floods	874
Queensland floods (QUFL)	ENG	29-Jan-2013	Floods	892
Alberta floods (ABFL)	ENG	19-Jun-2013	Floods	913
Manila floods (MNFL)	ENG	20-Aug-2013	Floods	808
Colorado floods (CLFL)	ENG	09-Sep-2013	Floods	901
Sardinia floods (SDFL)	IT + ENG	17-Nov-2013	Floods	910

Table 1. Crises datasets details, their types, and number of human tagged messages

3.3 Preprocessing

Preprocessing of the datasets is performed before running the experiments. Each crisis dataset is divided into two sets. The first set comprised of 70% of the messages (i.e. training set) and the second comprised of 30% of the messages (i.e. test set). For the both training and the test sets, we remove stop-words, URLs, and user mentions from the messages. We use two types of features uni-grams (one word) and bi-grams (two consecutive words). Feature selection is performed using the information gain feature selection method and top 1,000 features are selected for the training purposes. We use Random Forest, a well-known learning scheme (Liaw et al., 2002), as our classification algorithm. Results of all the experiments are presented in four well-known measures i.e. Precision, Recall, F-measure (i.e., weighted average F1), and AUC

(i.e. Area Under ROC curve).[4]

4. EXPERIMENTAL FRAMEWORK

To determine whether labeled data from past crises can be helpful in the classification of target crisis messages, we perform extensive experimentation.

In this paper, we always perform training on labeled data from one or more source events, and the trained models are always evaluated/tested on one target event. The evaluation set remains same for all types of experiments (more details below) for a given crisis event. The evaluation of models, especially in the domain adaptation setting, should be performed on a fixed test set, which is a more demanding evaluation task as compared to other types of evaluations such as cross-validation using *n*-folds. Hence, for an event under consideration, we fix 30% of its labeled data as test set for evaluations.

4.1 Model Performance for Event and In-domain settings

First, we perform a series of experiments to determine how models trained on same events perform as compared to when they are used to classify other events. For example, a model trained and tested on Nepal earthquake data vs. trained on Nepal and tested on Costa Rica earthquake.

4.2 Model Adaptation in In-domain and Cross-domain Settings

To test the performance of classifiers trained using labeled data from one event (source crisis) and test on another event (target crisis), we perform domain adaptation using single-source experiments. In this setting, we use datasets from both in-domain and cross-domains. The in-domain setting represents both train and test sets from same crisis type (e.g. earthquake). The cross-domain setting represents train and test from different crisis types.

In-domain (earthquakes)*:* First, we take earthquake datasets in their chronological order and use the event under investigation as target event and its preceding crises as source events. Although, considering chronological order of the events does not have any direct effect on machine learning or natural language processing techniques, we just want to mimic real-world setting. We always train classifiers on the source event data and test on the target event data. In Table 2, all the rows with experiment type "SS" (i.e., single-source) represent the results obtained using the single-source experiments. For instance, the first SS row in Table 2 shows the results of training on ITEQ 100% (i.e. all Italy earthquake labels) and testing on CREQ 30% (i.e. 30% of Costa Rica earthquake labels). The Italy earthquake event happened before the Costa Rica earthquake. And the reason why Italy EQ is not tested because we don't have any preceding event to this one.

In-domain (floods)*:* Next, the floods datasets are tested. As before, the current crisis data is considered as the target event and its preceding crises the source event(s). As always, we train classifiers on source event data and test on target event data. Table 3 shows the results of in-domain (floods) experiments in rows with experiment type as "SS".

[4] https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Cross-domain (earthquakes and floods)*:* In this setting, we performed cross-domain experiments i.e. both source and target datasets are taken from different domains. In these experiments, the aim is to find out if incorporating training examples from other crisis types can increase classification accuracy or not. Table 4 shows the results of cross-domain experiments for some selected events.

4.3 Model Adaptation Using Multiple Sources (in-domain)

To test whether incorporating more training examples from more than one similar past crises increases the classification accuracy or not, we perform the following two types of experiments.

1. Multiple sources without target event data

Obtaining labeled data during an ongoing event, especially in the early hours, is challenging and often not possible. In this experiment setting, we use past data only without any labeled data from a target event. One basic motivation behind this setting is the fact that more training examples tend to boost classifier's capability to generalize concepts better. To determine whether incorporating labels from all similar past crises is useful or not, in this experiment, we take all preceding datasets as our source events and used as training set. New models are trained using this training set. The evaluation of the newly generated models is performed on the test set of a target event.

Table 2 with rows having experiment type "MS" (i.e. multi-source) shows the results of all the earthquake events. Table 3 shows the results of all the floods events (rows with experiment type "MS").

2. Multiple sources with target event data

Given the fact that classifiers generalize better if both training and test instances are drawn from the same data distribution. In this setting, we include training examples from the target event. For this purpose, we take labels (70%) from the target event to determine the boost in classification accuracy. Table 2 shows the results of earthquake events and Table 3 shows the results of floods events, both with rows having experiment type as "MSWT" (i.e. multi-source with target event).

4.4. Model Adaptation in Cross-language and other Cases

In supervised classification systems that make use of textual features such as uni-grams, bi-grams, or part-of-speech tags, etc., the language of the underlying data from which the features are drawn play an important role. Two events of same type (e.g. earthquake) happened in two different countries could be effectively used to train classifiers, if the language of the both countries is similar (e.g., Italian and Spanish). To determine the usefulness of such cases, in this setting, we train and test classifiers in which both source and target events are from countries where the lexical similarity between their spoken languages is high. For instance, according to Wikipedia[5] the lexical similarity between Spanish and Italian language is almost 82%.

Rows with experiment type "SC" (i.e. special case) in Table 2 and Table 3 show the results of this analysis. For instance, in case of the Bohol Earthquake (BOEQ), we ran three additional tests. In the first test (SC1), we dropped ITEQ as it was present in the BOEQ MS case in which we observe a drop in the accuracy (e.g. see AUC).

[5] https://en.wikipedia.org/wiki/Lexical_similarity

However, after dropping the ITEQ, the classification accuracy increases (see SC1 row of BOEQ in Table 2). As the ITEQ set contains messages from both English and Italian languages, probably this causes the drop of AUC in the first test and the increase in AUC in the second test. To validate this observation, we manually analyzed all 912 ITEQ tweets to assign language tags (English or Italian). The result of the language tagging found that 90% of the tweets in ITEQ are in Italian language. Next, we only used ITEQ-EN (10% English set) along with CREQ and GUEQ to train a new model. The results are shown in the row with SC2 on BOEQ (30%) test set. We can see 9% increase in AUC.

Exp. Type	Source (s): Train set (size)	Target: Test set (size)	Precision	Recall	Weighted Avg. F1	AUC
SS	ITEQ (100%)	CREQ (30%)	0.76	0.56	0.57	0.85
MSWT	ITEQ (100%) + CREQ (70%)	CREQ (30%)	0.85	0.85	0.84	0.95
SS	CREQ (100%)	GUEQ (30%)	0.62	0.55	0.51	0.85
MS	ITEQ (100%) + CREQ (100%)	GUEQ (30%)	0.77	0.66	0.69	0.93
MSWT	ITEQ (100%) + CREQ (100%) + GUEQ (70%)	GUEQ (30%)	0.84	0.85	0.83	0.97
SS	GUEQ (100%)	BOEQ (30%)	0.73	0.42	0.48	0.73
MS	ITEQ (100%) + CREQ (100%) + GUEQ (100%)	BOEQ (30%)	0.76	0.49	0.55	0.68
MSWT	ITEQ (100%) + CREQ (100%) + GUEQ (100%) + BOEQ (70%)	BOEQ (30%)	0.90	0.87	0.87	0.95
SC1	CREQ (100%) + GUEQ (100%)	BOEQ (30%)	0.80	0.43	0.56	0.76
SC2	ITEQ-EN (100%) CREQ (100%) + GUEQ (100%)	BOEQ (30%)	0.77	0.45	0.51	0.77
SC3	ITEQ-EN (100%) + CREQ (100%) + GUEQ (100%) + BOEQ (70%)	BOEQ (30%)	0.88	0.85	0.85	0.97
SS	BOEQ (100%)	NEEQ (30%)	0.48	0.25	0.15	0.64
MS	ITEQ (100%) + CREQ (100%) + GUEQ (100%) + BOEQ (100%)	NEEQ (30%)	0.54	0.25	0.15	0.60
MSWT	ITEQ (100%) + CREQ (100%) + GUEQ (100%) + BOEQ (100%) + NEEQ (70%)	NEEQ (30%)	0.87	0.86	0.86	0.97
SC1	CREQ (100%) + GUEQ (100%) + BOEQ (100%)	NEEQ (30%)	0.53	0.29	0.21	0.63
SC2	ITEQ-EN (100%) + CREQ (100%) + GUEQ (100%) + BOEQ (100%) + NEEQ (70%)	NEEQ (30%)	0.86	0.86	0.86	0.98

Table 2. In-domain single-source (SS), multi-source (MS), multi-source with target crisis (MSWT), and special case (SC) model adaptation results for earthquake datasets

In the third test, we include 70% of the BOEQ labels along with ITEQ-EN, CREQ, and GUEQ.

For this the results can be seen in SC3 row of Table 2. When we use the ITEQ-EN, i.e., only the English language tweets related to the Italy earthquake, we noted an increase in the performance of new classifier.

Exp. type	Source (s): Train set (size)	Target: Test set (size)	Precision	Recall	Weighted Avg. F1	AUC
SS	PHFL (100%)	QUFL (30%)	0.60	0.50	0.51	0.82
MSWT	PHFL (100%) + QUFL (70%)	QUFL (30%)	0.86	0.85	0.85	0.97
SS	QUFL (100%)	ABFL (30%)	0.74	0.61	0.61	0.83
MS	PHFL (100%) + QUFL (100%)	ABFL (30%)	0.42	0.43	0.40	0.81
MSWT	PHFL (100%) + QUFL (100%) + ABFL (70%)	ABFL (30%)	0.80	0.80	0.79	0.96
SS	ABFL (100%)	MNFL (30%)	0.61	0.52	0.53	0.77
SC1	PHFL (100%)	MNFL (30%)	0.70	0.61	0.60	0.91
SC2	PHFL (100%) + MNFL (70%)	MNFL (30%)	0.77	0.75	0.75	0.95
MS	PHFL (100%) + QUFL (100%) + ABFL (100%)	MNFL (30%)	0.74	0.69	0.70	0.89
MSWT	PHFL (100%) + QUFL (100%) + ABFL (100%) + MNFL (70%)	MNFL (30%)	0.81	0.80	0.80	0.95
SS	MNFL (100%)	CLFL (30%)	0.65	0.54	0.48	0.85
SC	QUFL (100%) + ABFL (100%)	CLFL (30%)	0.75	0.67	0.70	0.94
MS	PHFL (100%) + QUFL (100%) + ABFL (100%) + MNFL (100%)	CLFL (30%)	0.80	0.76	0.76	0.94
MSWT	PHFL (100%) + QUFL (100%) + ABFL (100%) + MNFL (100%) + CLFL (70%)	CLFL (30%)	0.83	0.83	0.83	0.96
SS	CLFL (100%)	SDFL (30%)	0.55	0.41	0.29	0.78
MS	PHFL (100%) + QUFL (100%) + ABFL (100%) + MNFL (100%) + CLFL (100%)	SDFL (30%)	0.61	0.53	0.54	0.85
MSWT	PHFL (100%) + QUFL (100%) + ABFL (100%) + MNFL (100%) + CLFL (100%) + SDFL (70%)	SDFL (30%)	0.88	0.88	0.88	0.98

Table 3. In-domain single-source (SS), multi-source (MS), multi-source with target crisis (MSWT), and special case (SC) model adaptation results for floods datasets

For floods datasets, again rows with experiment type "SC" show the results of the special cases analysis. For instance, in case of MNFL, we can observe an increase in accuracy when using PHFL as train set as compared to PHFL, QUFL, and ABFL altogether for training (see rows "MS" and "SC1" of MNFL).

Source (s): Train set (size)	Target: Test set (size)	Precision	Recall	Weighted Avg. F1	AUC
BOEQ (100%)	PHFL (30%)	0.38	0.35	0.26	0.58
BOEQ (100%) + PHFL (70%)	PHFL (30%)	0.70	0.72	0.63	0.89
BOEQ (100%) + MNFL (100%)	PHFL (30%)	0.66	0.61	0.58	0.86
BOEQ (100%) + MNFL (100%) + ABFL (100%)	PHFL (30%)	0.64	0.61	0.58	0.86
NEEQ (100%)	MNFL (30%)	0.35	0.42	0.27	0.55
MNFL (100%)	NEEQ (30%)	0.43	0.31	0.25	0.59
SDFL (100%)	ITEQ (30%)	0.62	0.50	0.46	0.69
BOEQ (100%) + MNFL (100%)	NEEQ (30%)	0.50	0.28	0.22	0.64

Table 4. Cross-domain single-source model adaptation results for both earthquake and floods datasets

The general lesson learned from Table 2 is that including more training data, even from a mixed-language source, improves the accuracy significantly. However, the following are interesting observations.

Data from the Italy earthquake had a serious negative effect in some settings (Bohol earthquake and Nepal earthquake) but, it was useful in others (Costa Rica and Guatemala earthquakes). We believe that this exception is because 90% of the Italian earthquake data was in Italian, whereas our test case contained tweets related to earthquakes in Bohol and Nepal were primarily in English. This result seems to suggest that Italian is closer to Spanish as a language than English, an observation validated by multiple speakers of these languages and by the language-tree[6]. In cases where the language is significantly different, e.g., ITEQ versus BOEQ or NEEQ, it is better to leave the training set out. However, in these cases, it is best to select the training examples in ITEQ that are in English and using it to train in these cases, as we showed which increases the classifier performance.

A proposition could be made that we should segregate tweets based on language and use tweets from the same language to train and test. However, that is not an optimal strategy. Note that, for the target GUEQ, learning from the same language Costa Rica earthquake and testing it on GUEQ is worse than learning from combining the Spanish and Italian tweets. So, at least, when you do not have enough Spanish data to train, training using Italian was valuable and increased accuracy. Training data from target, even in small proportion, always help increase classifiers

[6] An interesting by-product of our work could be to construct a language-tree and language-language distances based on online language in disaster-related tweets. Perhaps such a tree/distance measure could be then used to select which languages can be used for cross-training and which should not be used, especially in cases where we have few training examples in one language.

performance. This can be seen in all experiments in which 70% of the target labels were included in the training set.

Table 3 also confirms the general philosophy that more training data is good. However, there are some interesting observations there too. For the test case MNFL, using the Philippines data and the Manila data performs almost as well as when we put the other data in. This shows that using data from the same area will be immensely useful because the language mixture (Tagalog, English) used in these two cases is almost exactly the same. However, adding QUFL and ABFL, which are solely in English, seems to improve the performance slightly. Colorado floods data does not help much on Sardinia floods, however, a combination of PHFL, QUFL, ABFL, MNFL, and CLFL improves the performance, which then reaches to an acceptable performance when SDFL is added in the training set.

Generally, we see that tagging a few tweets related to the same earthquake/flood still improves the performance significantly. Perhaps this may mean that we still do not have enough data and in the future, when we collect more data, we can eliminate the requirement for training on the current (target) dataset.

Initial results are promising to show that there may be some signal in using the flood related tweets to augment earthquake tweets but it also has the chance of reducing the accuracy of the classifier. For example, Table 4 (last row) shows that adding MNFL to BOEQ increased the performance on the test set NEEQ. However the previous two rows show a slight decrease in accuracy by adding the flood-related training set. The general consensus seems to be that given our collection of tagged tweets from the past, we should stick to using all the tweets from the same domain provided the language mixture is similar. At this point, using cross-domain training sets have not conclusively shown any consistent improvement in the accuracy.

As we have observed that the event-specific data always help improve classification performance of a classifier, we presented two alternative approaches to incorporate event-specific unlabeled data in the training phase. When there are no budget constraints, the approach that uses machine classified items with high confidence score from target domain are useful to be included in the training sets. However, under budget constraints, the active learning approach seems more feasible, as it achieves high accuracy with less labeled examples.

5. ENHANCING PERFORMANCE OF BASELINE MODELS

In the previous section, we show the utility of the labeled data from past crisis events using different settings. In this section, we consider the best performing models as baseline and try to enhance their classification performance using different techniques. Specifically, first we try to use target events unlabeled data to boost classifiers performance and second we employ active learning approach to determine if it helps improve classifier's performance.

5.1 Boosting Performance of Models Using Target Unlabeled Data

Obtaining event-specific labeled data to train machines in the early hours of a crisis situation is hard due to a number of reasons such as scarcity of human volunteers for manual annotation tasks. However, a vast amount of unlabeled event-specific data is available during those hours. In the above described experiments, we have observed that event-data plays a significant role in improving classification accuracy of a classifier when considered for machine training. For instance, in the cases of ABFL, MNFL, and SDFL, an increase of 16%, 6%, and 13%

respectively have been observed in AUC when event-labeled data included in training sets.

To test whether the unlabeled data of a target event can be useful or not when used as training set, in this section we perform a series of experiments. Specifically, we aim to determine whether target's unlabeled items, when classified using a pre-trained (using past events labeled data) classifier, used as training examples help improve classifier's performance or not. For this purpose, we take models from the above mentioned experiments, which outperform without event data, to automatically predict labels for a target's unlabeled items. We consider target's machine classified items for which the classification accuracy is high (e.g. > 80%) as potential candidates to be used as training examples along with existing labeled items from past events. Trained models are evaluated using the hold out test sets (30%) described in the previous sections.

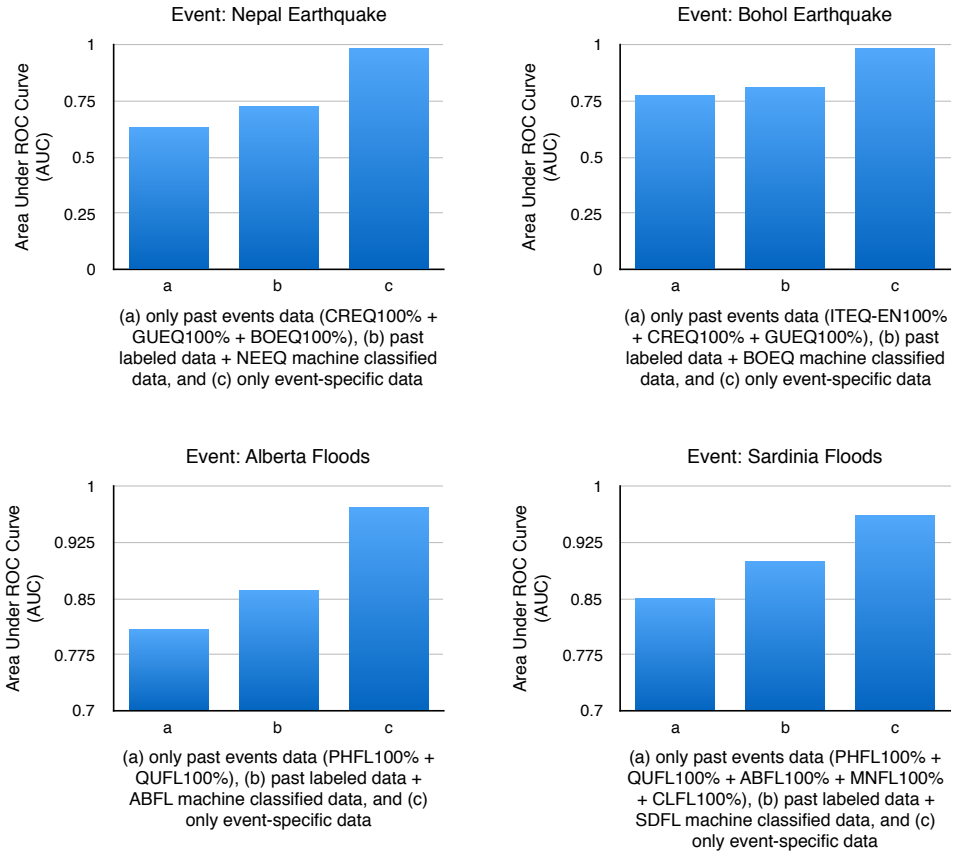

Figure 2: Boosting performance of classifiers using target's unlabeled data for four different events (two earthquakes and two floods)

Figure 2 depicts the results of our experiments on four different events. Specifically, we used two earthquakes and two floods events and tested three different strategies for training a classifier as follows:

(a) Using only past events labeled data as training examples

(b) Using machine labeled event-specific data as training examples

(c) Using only event-specific labeled data as training examples

We can see that in all four cases, the setting "b" is better than the setting "a". That is, when there

is no event-specific labeled data available in the early hours of a crisis event, it is feasible to take the unlabeled data from the event and use machines trained on past-labeled data for prediction. Then, select items for which the machine confidence is high. However, it is evident that the most the event-specific human-labeled data the better, as it can be clearly seen in the case of setting "c", in which we use human-labeled examples from the same event.

5.2 Boosting Performance of Models Using Active Learning

To enable the rapid classification of target events items, in this section the setting we consider for supervised learning assumes that we have a large amount of labeled data from past events (i.e. source domains), a large amount of unlabeled data from target domain (i.e. event itself), and a small budget in terms of money or availability of volunteers to help annotate messages in the early hours of a crisis event. Given this setting, the idea is to train a best performing classifier by intelligently utilizing the budget, only when it is necessary.

For this purpose, first we aim to learn a model using the past events labeled data and then employ an active learning technique to improve the model's performance. Active learning is a well-known approach in many machine learning problems where unlabeled data is abundant, but labeled data is scarce or very expensive to obtain (Settles 2010; Cohn et al. 1994). Basically, once employed, the active learning approach in a supervised learning system overcome the issue of scarcity of labeled data by selecting unlabeled items from the target event to be labeled by a human annotator. However, a number of different strategies can used to decide which items from target domain should be selected for labeling. In this work, we follow the one in which the learner selects the items from the target unlabeled data about which the model is least certain about its label.

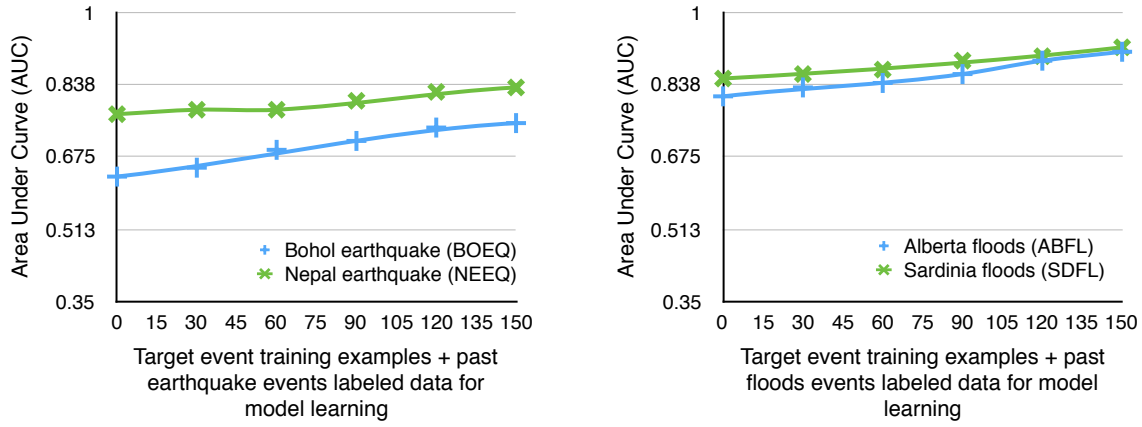

Figure 3: Performance improvement of classifiers using active learning approach for four events (two earthquake and two floods)

Specifically, we begin with a model trained on the labeled data from past events. Next, the active learning method picks items for which it is unsure about their classification, for example, the items that are close to the decision boundary and for which the labels are maximally informative. Human annotators tag the selected target event's items. Finally, the human annotated items are included in the training set of the classifier. Figure 3 shows the results obtained using the active learning approach on both earthquakes and floods events. We used best performing models on past events data and used active learning to select the items which could be more beneficial for the baseline models. A model is trained after receiving a batch of 30 labeled items from target

event and evaluated using the hold out test sets (30%) described in the previous sections.

The results clearly show the utility of the active learning approach as we can see that models achieve reasonable accuracies with fewer labeled examples.

6. RELATED WORK

Mass convergence and disaster events, particularly those with no prior warning, require rapid analysis of the available information to make timely decisions (Castillo 2016; Imran et al., 2015). With the proliferation of the Web 2.0 technologies, handheld devices, and other sensors, a number of opportunities have emerged ranging from early detection of a disaster to perform extensive aftermath analysis (Palen 2008; Brownstein et al., 2009). Many research studies analyzed the usefulness of this online information for humanitarian response organizations (Kavanaugh et al., 2012; Palen et al., 2010; Vieweg et al., 2014; Dashti et al., 2013). Starbird et al. (Starbird et al., 2010) analyzed microblog usage and information lifecycles during disaster situations. The authors found that the information including geo-location, situational updates and warnings can contribute to situation awareness and are typically communicated during each crisis-incident on Twitter. Hughes and Palen (Hughes et al., 2009) examined the use of Twitter during four high profile disaster events and emergencies. The authors observed that tweets posted during such crisis events reveal features of information that can support information broadcasting and brokerage.

Despite a number of issues regarding the trustworthiness and credibility of social media information have been identified (Tapia et al., 2011; Castillo et al., 2011; Morris et al., 2012), many studies found information posted on microblogging platforms during crises can aid crisis response efforts, if processed timely and correctly (Yin et al. 2012; Starbird et al., 2010; Palen et al., 2009). Many approaches based on human annotation, supervised learning, and unsupervised learning techniques have been proposed to process social media data---for a complete survey see e.g., (Imran et al., 2015).

In this work, we use supervised machine learning techniques to classify crisis-related messages. Many such efforts and systems based on machine learning techniques have been developed in past e.g., (Mendoza et al., 2010; Olteanu 2015; MacEachren et al., 2011; Imran et al., 2014; Roy et al., 2013). For instance, ESA (Yin et al. 2012; Power et al. 2014) uses naıve Bayes and SVM, EMERSE (Caragea et al., 2011) uses SVM, AIDR (Imran et al. 2014) uses random forests, and Tweedr (Ashktorab et al. 2014) uses logistic regression. The basic assumption for all these systems to achieve high classification accuracy is the availability of fairly big human-labeled data. However, due to the scarcity of training data, which is one of the basic ingredients for such approaches to work well, causes delays in machine training thus humanitarian efforts in time-critical situations cannot be launched effectively.

To overcome the issue of the scarcity of training data during a new crisis and emergency, we study the usefulness of human-labeled data (training data) from past crises. Li et al., (2015) studied the problem of domain adaptation. They combine source labeled data with target unlabeled data to train classifiers (Naive Bayes in their case) and observed a high performance by including target crisis data in training set as compared to only source crisis data. Their findings, to some extent, are inline with ours; however, the evaluation mechanism that they have used is based on cross-validation using 5-fold setting. However, in our case, we always use a holdout test set across all variations of experiments, which is a more challenging problem in an

online classification setting. Moreover, we also provide empirical results by training models in cross-language settings.

7. CONCLUSIONS

Availability of training data to train machine learning classifiers during the early hours of a crisis situation can help gain early insights for rapid crisis response. We show that using labeled data from past events of the same type are generally always useful if the training and testing data are from the same language. When there are not enough tweets in the one language (e.g., Spanish), labeled tweets in a different language (e.g., Italian) can be useful if the two languages in question are very similar (e.g., Italian and Spanish) but not when they are not (e.g., Italian and English/Tagalog). If there are reasonable number of labeled tweets from the same domain (e.g., earthquakes), then, we could not establish the utility of using labeled tweets from a different domain (e.g., floods). In one such case, the performance improved slightly while in another it decreased. Overall, event-specific labeled data always help to boost models performance. For this reason, we also presented two approaches to improve the performance of a baseline model. Both approaches, depending on budget constraints, help improve classification performance.

REFERENCES

1. Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. Proc. of ISCRAM.
2. Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection—harnessing the Web for public health surveillance. New England Journal of Medicine, 360(21), 2153-2157.
3. Cameron, M. A., Power. A., Robinson, B., and Yin, J. (2012). Emergency Situation Awareness from Twitter for Crisis Management. In Proc. Conference on World Wide Web (WWW).
4. Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H. W., Mitra, P., Wu, D., Tapia, A., Giles, L., Jansen, B., & Yen, J. (2011, May). Classifying text messages for the haiti earthquake. In *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM 2011)*.
5. Castillo C. Big Crisis Data. Cambridge University Press; 2016 Jul 4.
6. Castillo, C., Mendoza, M., & Poblete, B. (2011, March). Information credibility on twitter. In Proceedings of the 20th international conference on World wide web (pp. 675-684). ACM.
7. Cohn, D., Atlas, L., & Ladner, R. (1994). Improving generalization with active learning. Machine learning, 15(2), 201-221.
8. Daume III, H., & Marcu, D. (2006). Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 101-126.
9. Dashti, S., Palen, L., Heris, M. P., Anderson, K. M., Anderson, S., & Anderson, S. (2014, May). Supporting disaster reconnaissance with social media data: a design-oriented case study of the 2013 Colorado floods. In Proceedings of the 11th International ISCRAM Conference (pp. 18-21).

10. Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, *6*(3-4), 248-260.
11. Imran, M., Elbassuoni, S. M., Castillo, C., Diaz, F., & Meier, P. (2013a). Extracting information nuggets from disaster-related messages in social media. *Proc. of ISCRAM, Baden-Baden, Germany*.
12. Imran, M., Lykourentzou, I., Naudet, Y., & Castillo, C. (2013b). Engineering crowdsourced stream processing systems. arXiv preprint arXiv:1310.5463.
13. Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). AIDR: Artificial intelligence for disaster response. In *Proceedings of the companion publication of the 23rd international conference on World Wide Web companion* (pp. 159-162). International World Wide Web Conferences Steering Committee.
14. Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, *47*(4), 67.
15. Kavanaugh, A. L., Fox, E. A., Sheetz, S. D., Yang, S., Li, L. T., Shoemaker, D. J., ... & Xie, L. (2012). Social media use by government: From the routine to the critical. Government Information Quarterly, 29(4), 480-491.
16. Liaw, A., & Wiener, M. (2002). Classification and regression by Random Forest. *R news*, *2*(3), 18-22.
17. Li, Hongmin, Nicolais Guevara, Nic Herndon, Doina Caragea, Kishore Neppalli, Cornelia Caragea, Anna Squicciarini, and Andrea H. Tapia. (2015). Twitter Mining for Disaster Response: A Domain Adaptation Approach. ISCRAM 2015.
18. Mendoza, M., Poblete, B., & Castillo, C. (2010, July). Twitter Under Crisis: Can we trust what we RT?. In *Proceedings of the first workshop on social media analytics* (pp. 71-79). ACM.
19. MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., & Blanford, J. (2011, October). Senseplace2: Geotwitter analytics support for situational awareness. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (pp. 181-190). IEEE.
20. Morris, M. R., Counts, S., Roseway, A., Hoff, A., & Schwarz, J. (2012, February). Tweeting is believing?: understanding microblog credibility perceptions. In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (pp. 441-450). ACM.
21. Neubig, G., Matsubayashi, Y., Hagiwara, M. and Murakami, K. (2011). Safety information mining – what can NLP do in a disaster. In Proc. International Joint Conference on Natural Language Processing (IJCNLP)
22. Olteanu, A., Vieweg, S., & Castillo, C. (2015, February). What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of CSCW,* (pp. 994-1009). ACM.
23. Palen L. Online social media in crisis events. Educause Quarterly. 2008 Jul;31(3):76-8.
24. Palen, L., Vieweg, S., Liu, S. B., & Hughes, A. L. (2009). Crisis in a networked world features of computer-mediated communication in the April 16, 2007, Virginia Tech Event. *Social Science Computer Review*, *27*(4).

25. Palen, L., Starbird, K., Vieweg, S., & Hughes, A. (2010). Twitter-based information distribution during the 2009 Red River Valley flood threat. Bulletin of the American Society for Information Science and Technology, 36(5), 13-17.
26. Power, R., Robinson, B., Colton, J., & Cameron, M. (2014). Emergency situation awareness: Twitter case studies. In *Information Systems for Crisis Response and Management in Mediterranean Countries* (pp. 218-231).
27. Roy Chowdhury, S., Imran, M., Asghar, M. R., Amer-Yahia, S., & Castillo, C. (2013). Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In *10th International ISCRAM Conference*, *Baden-Baden, Germany*.
28. Sakaki, T., Okazaki, M. and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proc. World Wide Web Conference (WWW).
29. Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010, February). Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 241-250). ACM.
30. Settles, B. (2010). Active learning literature survey. University of Wisconsin, Madison, 52(55-66), 11.
31. Tapia, A. H., Bajpai, K., Jansen, B. J., Yen, J., & Giles, L. (2011, May). Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In Proceedings of the 8th International ISCRAM Conference (pp. 1-10).
32. Tapia, A. H., Moore, K. A., & Johnson, N. J. (2013, May). Beyond the trustworthy tweet: A deeper understanding of microblogged data use by disaster response and humanitarian relief organizations. In Proceedings of the 10th International ISCRAM Conference (pp. 770-778). Baden-Baden.
33. Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. *In Proc. SIGCHI,* (pp. 1079-1088). ACM.
34. Vieweg, S., Castillo, C., & Imran, M. (2014). Integrating social media communications into the rapid assessment of sudden onset disasters. In Social Informatics (pp. 444-461). Springer International Publishing.
35. Qu, Y., Huang, C., Zhang, P., and Zhang, J. (2011). Microblogging After a Major Disaster in China: A Case Study of the 2010 Yushu Earthquake. In Proc. CSCW.
36. Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, (6), 52-59.