Disaster Response Aided by Tweet Classification with a Domain Adaptation Approach

Hongmin Li[1], Doina Caragea[1], Cornelia Caragea[2], Nic Herndon[3]
[1]Department of Computer Science, Kansas State University, USA
[2]Department of Computer Science and Engineering, University of North Texas, USA
[3]Department of Ecology and Evolutionary Biology, University of Connecticut, USA
{hongminli, dcaragea}@ksu.edu, ccaragea@unt.edu, dnic.herndon@uconn.edu

Abstract

Social media platforms such as Twitter provide valuable information for aiding disaster response during emergency events. Machine learning could be used to identify such information. However, supervised learning algorithms rely on labeled data, which is not readily available for an emerging target disaster. While labeled data might be available for a prior source disaster, supervised classifiers learned only from the source disaster may not perform well on the target disaster, as each event has unique characteristics (e.g., type, location, culture) and may cause different social media responses. To address this limitation, we propose to use a domain adaptation approach, which learns classifiers from unlabeled target data, in addition to source labeled data. Our approach uses the Naïve Bayes classifier, together with an iterative Self-Training strategy. Experimental results on the task of identifying tweets relevant to a disaster of interest show that the domain adaptation classifiers are better as compared to the supervised classifiers learned only from labeled source data.

Keywords: *Twitter, Domain Adaptation, Classification, Disaster Response*

1 Introduction

Nowadays, social media platforms, such as Twitter, are widely used for sharing and spreading information and news during emergency events. First-hand information produced by people in the affected areas is especially valuable as such information cannot be easily obtained from other sources (Landwehr & Carley, 2014). Furthermore, it has been suggested that people are very motivated to help and offer support to victims during emergency events (either natural disasters or man-made disasters), and such behaviors extend to social network users as well (Kaufhold & Reuter, 2016; Palen & Vieweg, 2008). As an example, during the Paris attacks in November 2015, eyewitnesses, or friends of eyewitnesses, shared information about gunfire and dangerous places through Twitter, to alert people within minutes after attacks in different places (BBC Trending, 2015). Parisians also launched the hashtag #PorteOuverte (meaning “open door”) to offer, through Twitter, safety and refuge to those affected by the attacks (Murdock, 2015). Therefore, microblogging data from Twitter-like platforms are seen to have intrinsic value for both responder organizations and victims, due to their growing ubiquity, communications rapidity, and cross-platform accessibility (Vieweg *et al.*, 2010). Governments that aim to improve situation awareness during emergencies are also starting to value such on-the-ground information offered by eyewitnesses or average citizens on social media during emergency event (Homeland Security, 2014; Hughes *et al.*, 2014;Palen *et al.*, 2009; Reuter *et al.*, 2015).

There are numerous challenges when considering the use of social media data for emergency response, including issues of reliability, quantification of performance, deception, focus of attention, and translation of reported observations into a form that can be used to combine with other information. One problem became apparent during the earthquake in Haiti when thousands of technical volunteers from around the world suddenly attempted to provide responders with mapping capabilities, translation services, people and resource allocation, all via SMS at a distance (Harvard Humanitarian Initiative, 2011; Meier, 2013). As Meier (2013) stated: “We quickly realized that our platform was not equipped to handle this high volume and velocity of urgent information.” Despite the good will of field staff, their institutions' policies and procedures were never designed to incorporate data from outside their networks, especially at such an overwhelming flow. In addition, the organizations did not have the technical staff, or the analytical tools, to turn the flow of data into actionable knowledge (Harvard Humanitarian Initiative, 2011; Palen *et al.*, 2010; Tapia & Moore, 2014). Still, researchers have been optimistic about the potential value of microblogging data in helping emergency teams to improve situational awareness and emergency response, provided that accurate information can be automatically identified (Castillo, 2016; Meier, 2015; Palen & Anderson 2016; Qadir *et al.*, 2016; Reuter *et al.*, 2015; Watson *et al.*, 2017).

Temporally, the above problems arise at the stage when emergency responders and organizations begin engaging their organizational mechanisms to respond to the crises in question (Munro, 2011). For decades, these organizations have operated with a centralized command structure, standard operating procedures, and internal vetting standards to ascertain appropriate responses to emergencies. While not optimized to current expectations of speed, efficiency and knowledge, these mechanisms have been successful at bringing rescue, response and recovery to millions (Dugdale *et al.*, 2012; Walton *et al.*, 2011).

Towards optimizing current organizational mechanisms in terms of speed, efficiency and knowledge, supervised machine learning algorithms have been used to help responders sift through the big crisis data, and prioritize information that may be useful for response and relief (Ashktorab *et al.*, 2014; Beigi *et al.*, 2016; Caragea *et al.*, 2014; Gao *et al.*, 2011; Imran *et al.*, 2015; Kumar *et al.*, 2014; Terpstra *et al.*, 2012; Yin *et al.*, 2012). Supervised algorithms require labeled training data to learn classifiers that can be further used to label more data from the same domain. The labels are precisely the categories that we want to associate with data of interest, for example *relevant* or *non-relevant* for tweets relevant to a disaster, *positive* or *negative* in the case of tweets that express a sentiment, or even the specific sentiment expressed by a tweet (*anger*, *compassion*, *sorrow*, *fear*, etc.). Labeling of the training data is usually done manually, and is therefore a time-consuming and expensive process. This presents a big challenge when trying to use supervised learning algorithms to aid disaster response in the case of a new disaster, as the time and effort required to label tweets from that disaster prevents the timely usage of the classifiers.

To address this challenge, several works proposed to use labeled data from prior “source” disaster to learn *supervised* classifiers for a *target* crisis (Caragea *et al.*, 2016; Imran *et al.*, 2016; Verma *et al.*, 2011).

One drawback of this approach is that supervised classifiers learned in one crisis event, do not generalize well to other events (Imran *et al*., 2015; Qadir *et al*., 2016), as each event has unique characteristics (e.g., type, location, culture) and may cause different social media responses (Palen & Anderson 2016). As suggested by Li *et al*. (2015), this is particularly true for classification tasks that aim to identify tweets relevant to a disaster of interest, as features/words specific to that disaster represent important cues for the classification task. Unsupervised domain adaptation approaches (Jiang, 2008; Pan & Yang, 2010) that make use of unlabeled data from the target crisis in addition to label data from source crises are desirable.

A systematic literature review showed that (Li *et al*., 2015) was the first work that applied domain adaptation algorithms in the context of disaster response. The domain adaptation algorithm proposed by Li *et al*. (2015) was based on an iterative Expectation-Maximization (EM) approach, where a classifier is learned at each iteration, and used to assign soft (probabilistic) labels to the target unlabeled data. Subsequently, the target data is used to train the classifier at the next iteration. Experimental results on a small dataset from the Hurricane Sandy and Boston Marathon Bombings suggested that the domain adaptation approach that makes use of target unlabeled data, in addition to source labeled data, is better than supervised learning approaches that only use of source labeled data. This is especially true for domain specific tasks, such as the task of identifying tweets that are relevant to a disaster of interest.

Similar to Li *et al.* (2015), we use a domain adaptation approach to address the problem of identifying tweets that are relevant to a disaster of interest, among all the tweets that are posted during that disaster. Given the availability of a large set of crisis event tweets published by Olteanu *et al.* (2014), our goal is to extend the study of the domain adaptation approach proposed by Li *et al*. (2015) to gain more insights into its behavior when presented with a larger number of different types of disasters and larger amount of data for each disaster. As opposed to Li *et al*. (2015), who used EM with soft-labels when adding target data to the training set during the domain adaptation iterations, we propose the use of self-training with hard-labels as a prior text classification study (Nigam & Ghani, 2000) suggests that self-training with hard-labels can give better results than EM with soft-labels. Our main contributions are summarized as follows:

- We propose a modified version of the weighted Naïve Bayes domain adaptation algorithm introduced by Li *et al*. (2015). Our version uses the self-training strategy with hard-labels, instead of EM with soft-labels, to identify disaster related tweets.
- We perform experiments with a large dataset collected from several disasters (Olteanu *et al*., 2014) to better understand how much a domain adaptation algorithm can help improve a ready-to-use classifier trained only on labeled data from a previous disaster.
- We compare our self-training based classifiers with both supervised Naïve Bayes classifiers learned from source only and EM based classifiers learned using the original approach proposed by Li *et al.* (2015).

2 Related Work

Domain adaptation methods (Blum & Mitchell, 1998; Yarowsky, 1995) are not new in machine learning, and have been extensively used in areas such as text classification (Dai *et al*. 2007), sentiment analysis (Tan *et al.* 2009), bioinformatics (Herndon & Caragea, 2014; 2015). In what follows, we will review works that are closely related to our approach and application. For more details on domain adaptation approaches and applications, the reader is referred to the surveys by Jiang (2008) and Pan & Yang (2010).

Dai *et al.* (2007) proposed a domain adaptation algorithm, based on the Naïve Bayes classifier and EM, to classify text documents from Newsgroups, SRAA, and Reuters into top categories. Experimental results showed that this algorithm performs better than supervised algorithms based on either Support Vector Machine (SVM) or Naïve Bayes classifiers.

Tan *et al*. (2009) proposed a weighted version of the multinomial Naïve Bayes classifier combined with EM for sentiment analysis. Their algorithm filters out domain specific features from the source domain, by keeping only the top-ranking features that have similar probabilities in both source and target domains. In the first step, a Naïve Bayes classifier is trained on the source data and used to label the unlabeled data from the target domain. In subsequent iteration, the EM algorithm is used with a weighted combination of the source and target data to train a new Naïve Bayes classifier. Specifically, in the maximization (M) step, the prior and likelihood are calculated, and in the expectation (E) step, the posterior is calculated for the instances in the target data. These steps are repeated until convergence, with the weight shifting from the source to the target domain, iteration by iteration.

Herndon and Caragea (2015) proposed an approach like the one in (Tan *et al*., 2009) for the task of splice site prediction. They used a weighted Naïve Bayes classifier, and three methods for incorporating the target unlabeled data: EM with soft-labels, ST with hard-labels, and also a combination of EM/ST (with hard-labels for the most confidently labeled instances in the current target unlabeled data, and soft-labels for the other instances). They found that for the task of splice site prediction, EM with soft-labels gives better classifiers than the other two methods, contrary to what has been observed on text classification (Nigam & Ghani, 2000), where ST with hard-labels gives better results.

Peddinti & Chintalapoodi (2011) used domain adaptation to perform sentiment classification of tweets. Given a source dataset, in addition to target labeled data, they proposed two methods to identify source instances that can improve the classifier for the target, based on EM and Rocchio SVM. Namely, at each EM iteration, they first used target labeled data to classify source instances, then selected the most confident source instances to add back to the training set. Therefore, this method requires a small amount of target labeled data.

In the context of disaster response and rescue, there are several studies that applied machine learning and natural language processing (NLP) methods for disaster management (Kumar *et al*., 2014; Purohit *et al*., 2013; Sakaki *et al*., 2010; Terpstra, 2012). In particular, several works have studied supervised learning algorithms in regard to transferring information from a prior source disaster to a current target disaster (Imran *et al*., 2013a; 2016; Verma *et al.,* 2011), and will be reviewed in what follows.

Verma *et al.* (2011) used natural language processing techniques together with machine learning algorithms, Naïve Bayes and Maximum Entropy, to identify situational awareness tweets during crisis events. They used data from four crisis events, Red River Flood in 2009 and 2010, Haiti Earthquake in 2010 and Oklahoma Grass Fire in 2009. They first built two supervised classifiers with Naïve Bayes and Maximum Entropy to classify situation awareness tweets from each of the four crisis events, respectively. Subsequently, they also studied how well the classifiers performed across the four events. They found that the classifiers generalized well across the Red River Flood 2009 and Red River Flood 2010 events, but not for other events. For example, the performance (measured as accuracy) was poor when using the classifier learned from the Haiti Earthquake data to classify the Oklahoma Grass Fire data and vice versa, because these two types of events differ from each other in many aspects.

Imran *et al.* (2013a) performed similar experiments with two disasters, namely Joplin Tornado (as source) and Hurricane Sandy (as target), to identify information nuggets using conditional random fields (CRF). After classifying different types of informative (*casualties*, *donations*, etc.) tweets with Naïve Bayes classifiers, they used a sequence labeling algorithm, conditional random fields (CRF), to extract useful information, such as the number of casualties or the name of the infrastructure. They learned supervised classifiers either from source, or from source and 10% of labeled target data. They tested these classifiers on all target data and remaining 90% of target data, respectively, and compared the domain adaptation results with the results of supervised classifiers learned from 66% of labeled target data, and tested on 33% target data. Their experiments showed that using source data only results in a significant drop in the detection rate, while not affecting significantly the recall.

Imran *et al.* (2016) studied the usefulness of previous disaster tweets, and also the usefulness of using data in different languages. They experimented with several pairs of disasters, earthquakes and floods, from different countries. They learned a Random Forest classifier from a source disaster to classify a target disaster. Their results also showed that data from prior disasters of the same type as the current disaster can be very useful even across different languages.

While these works represent great steps towards using domain adaptation for disaster and crisis situations, the performance of the supervised classifiers used across different types of disasters or events is still poor, especially for domain specific tasks (e.g., identifying tweets relevant to a certain disaster). Domain adaptation techniques that have been successfully used in text classification, sentiment analysis, etc. hold great promise for classification problems in disaster and crisis management as well.

Inspired by the success of previous domain adaptation approaches, Li *et al*. (2015) proposed a domain adaptation algorithm, which made use of unlabeled data from the target disaster, together with labeled data from the source disaster. Li *et al*. (2015) studied their proposed domain adaptation algorithm on three classification tasks from two disasters, Hurricane Sandy (used as source) and Boston Marathon Bombing (used as target), with promising results especially on the task of identifying tweets relevant to a certain disaster. However, the algorithm was not tested extensively given that only data from two disasters was available, and the amount of data from each disaster was limited. Therefore, the main goal of our work is to extend the work by Li *et al*. (2015) and perform an extensive study of EM and ST domain adaptation approaches in the context of identifying tweets for a disaster of interest.

3 Methods

In this section, we will first introduce the supervised Naïve Bayes classifier, which is used as base classifier in our approach, and then describe our domain adaptation approach.

3.1 Naive Bayes Classifier

The Naïve Bayes algorithm is part of a class of algorithms known as Bayesian classification algorithms (Dai *et al*. 2007; Lewis, 1992; Manning *et al*., 2008). In particular, our approach uses a Naïve Bayes algorithm that is based on a multivariate Bernoulli model (Manning *et al*., 2008). Given a collection of documents $\boldsymbol{D}$ as training set, each document $\boldsymbol{d_i} \in \boldsymbol{D}, (i = 1, \ldots, N)$ represents a data instance, and has a class label $c_k \in \boldsymbol{C}$ associated with it. The set of words w_t in the collection $\boldsymbol{D}$ corresponds to the set of features used to represent documents, a.k.a., vocabulary $\boldsymbol{V}$. Using the features in $\boldsymbol{V},$ each document $\boldsymbol{d_i}$ is represented as a $|\boldsymbol{V}|$ dimensional vector of 0s and 1s, based on the occurrence of word $w_t \in \boldsymbol{V}$ in document $\boldsymbol{d_i}$. Using the Bayes rule and the assumption that features are independent given the class, the class label for a new document $\boldsymbol{d}$ can be obtained as:

$$c^* = \underset{c_k}{\operatorname{argmax}} P(c_k|\boldsymbol{d}) = \underset{c_k}{\operatorname{argmax}} \frac{P(\boldsymbol{d}|c_k)P(c_k)}{P(\boldsymbol{d})} = \underset{c_k}{\operatorname{argmax}} P(c_k) \prod_{t=1}^{|V|} P(w_t|c_k)$$

Therefore, to be able to predict the class label for new documents $\boldsymbol{d}$, we need to estimate the prior class probabilities $P(c_k)$ for all $c_k \in \boldsymbol{C}$, and the likelihoods $P(w_t|c_k)$ for all $w_t \in \boldsymbol{V}$ and $c_k \in \boldsymbol{C}$. Estimation of the class priors and likelihoods can be done based on a training data, and the process is generally referred to as training the Naïve Bayes classifier. Specifically, we estimate the class priors and likelihoods from the training data, using the add-1 smoothing strategy (to avoid zero probabilities), as follows:

$$P(c_k) = \frac{N(c_k) + 1}{N + 1}$$

$$P(w_t = 0 \mid c_k) = \frac{N(w_t = 0, c_k) + 1}{N(c_k) + 2}$$

$$P(w_t = 1 \mid c_k) = \frac{N(w_t = 1, c_k) + 1}{N(c_k) + 2}$$

where N is the total number of documents in the collection $\boldsymbol{D}$, $N(c_k)$ is the number of documents in class c_k, $N(w_t = 0, c_k)$ is the number of documents in class c_k that don't contain the word w_t, and $N(w_t = 1, c_k)$ is the number of documents in class c_k that contain the word w_t.

3.2 Domain Adaptation with Naïve Bayes Classifier

We will use a domain adaptation algorithm to identify tweets related to the current disaster. Our assumption is that there are no labeled tweets for the current target disaster, although unlabeled tweets are quickly accumulating. Furthermore, there exists a prior source disaster with labeled tweets. The goal is to use a domain adaptation approach that can make use of both source labeled data and target unlabeled data, while learning a classifier for the target.

Our approach is an adaptation of the iterative Expectation-Maximization (EM) approach used by Li *et al.* (2015). In the EM approach, a classifier is learned at each iteration, and used to label the target unlabeled data. Subsequently, the target data, with probabilistic soft-labels assigned by the current classifier (e.g., *p*(+|*d*)=0.7 and *p*(-|*d*)=0.3 for an instance *d*), are combined with the labeled source data and used to train the classifier at the next iteration. The original classifier is trained from source data only. The process continues for a fixed number of iterations, or until convergence.

Similar to the EM strategy, our proposed self-training approach is also an iterative approach that uses a weighted Naïve Bayes classifier to combine source and target data. As the EM approach, it starts by learning a supervised classifier from source data only, and uses that classifier to label the target unlabeled data. However, instead of adding all the target data with probabilistic soft-labels to the training set for the next iteration as in EM, in self-training, only the most confidently classified data are added to the training set, with hard (e.g., +/- or 1/0) labels.

More precisely, only the most confidently labeled instances (e.g., the top *k* instances in each class based on the posterior class distribution, *p(*+|*d)* and *p(*-|*d),* respectively) are added to the training set at subsequent iterations, and a new classifier is learned from the added target instances together with the original source instances. Given that we use the most confidently labeled target instances, which presumably have high posterior class distributions, we incorporate these instances with hard-labels, as opposed to probabilistic soft-labels, to help keep a cleaner decision boundary between the two classes.

The details of the algorithm are shown in Table 1. We denote the training source labeled data by tSL and the training target unlabeled data by tTU.

Table 1. Pseudocode for the domain adaptation algorithm

```
1. Simultaneously estimate the priors and likelihoods (a.k.a., train a Naïve Bayes classifier) for the
   source domain:
       $P(c_k) = P_{tSL}(c_k)$
       $P(w_t | c_k) = P_{tSL}(w_t | c_k)$
2. Use the classifier learned from source to assign labels to the unlabeled instances from
   the target domain, and select the most confidently labeled instances (based on the prior class
   distribution, e.g., top 5 instances in each class, for a balanced dataset) as hard-labeled
   instances for self-training.
3. Loop until the labels assigned to the remaining unlabeled target instances don't change:
    a) M-step: Same as Step 1, but use also the target instances labeled so far:
           $P(c_k) = (1-\gamma) \cdot P_{tSL}(c_k) + \gamma \cdot P_{tTU}(c_k)$
           $P(w_t | c_k) = (1-\gamma) \cdot P_{tSL}(w_t | c_k) + \gamma \cdot P_{tTU}(w_t | c_k)$
       where $\gamma = min\ (\tau\delta, 1)$, $\tau$ is the iteration number, and $\delta$ is a parameter that
       determines how fast we shift the weight from the source labeled data used for
       training (tSL) to the (originally unlabeled) target data used for training (tTU).
    b) E-step: Calculate the posterior class distribution for the remaining set of unlabeled
       instances from the target domain.
4. Use the final classifier to label test target unlabeled instances $\boldsymbol{d} = (w_1, \dots, w_{|V|})$:
       $c^* = \underset{c_k}{\operatorname{argmax}} P(c_k) \prod_{t=1}^{|V|} P(w_t|c_k)$
   where $V$ is the set of features/words used to represent instances.
```

4 Dataset and Preprocessing

We use a dataset of tweets, called CrisisLexT6, published by Olteanu *et al.* (2014). The tweets in the dataset are collected through Twitter API based on keywords and geo-locations of affected areas, and manually labeled as relevant to a disaster (on-topic) or not (off-topic) through the crowdsourcing platform Crowdflower. As discussed in the introduction, for a new disaster, the task of identifying tweets relevant to that disaster (on-topic), among all the tweets posted during the disaster, is the first task that needs to be addressed. Furthermore, this task is particularly suitable for domain adaptation, which uses prior source labeled data together with target unlabeled data, and can thus capture specific patterns in the target data itself.

The CrisisLexT6 dataset consists of six disasters occurring between October 2012 and July 2013 in USA, Canada and Australia. There are approximately 10,000 labeled tweets for each disaster. Similar to Li *et al.* (2015), we use the bag-of-words 0/1 representation to represent a tweet as a vector of features. We also use the same cleaning steps: removing non-printable ASCII characters; replacing URLs, email addresses, and usernames with an

URL/EMAIL/USERNAME placeholder, removing RT (*i.e.*, re-tweets) and duplicate tweets etc. The statistics for the final dataset are shown in Table 2, organized based on the time when each disaster happened.

Table 2. Statistics for the dataset used before cleaning and after cleaning

	Before Cleaning			**After Cleaning**		
Crisis	**On-topic**	**Off-topic**	**Total**	**On-topic**	**Off-topic**	**Total**
2012_Sandy_Hurricane	6,138	3,870	10,008	5,261	3,752	9,013
2013_Queensland_Floods	5,414	4,619	10,033	3,236	4,550	7,786
2013_Boston_Bombings	5,648	4,364	10,012	4,441	4,309	8,750
2013_West_Texas_Explosion	5,246	4,760	10,006	4,123	4,733	8,856
2013_Oklahoma_Tornado	4,827	5,165	9,992	3,209	5,049	8,258
2013_Alberta_Floods	5,189	4,842	10,031	3,497	4,714	8,211

From Table 2, we can see that the CrisisLexT6 disaster datasets are fairly balanced (i.e., the ratio of on-topic to off-topic tweets is close to 1). Following the timeline, in our experiments, we select pairs of source and target disasters, such that the source disaster happens first, while the target disaster happens at a later time.

5 Experimental Setup

Our goal is to evaluate the domain adaptation approach for the task of identifying tweets relevant to a target disaster (i.e., on-topic versus off-topic tweets). Our main working hypothesis is that the domain adaptation approach, which makes use of target unlabeled data in addition to source labeled data, can better capture patterns specific to the target as compared to a supervised learning approach that would use only source labeled data. To verify this hypothesis, we ask the following questions:

- How do supervised classifiers learned only from source labeled data perform on target data?
- How do the results of the domain adaptation classifiers, which use both source labeled data and target unlabeled data, compare with the results of the supervised classifiers, which use only source data, when used to classify target data?

Given that our domain adaptation approach uses self-training with hard-labeled target data, as opposed to EM with soft-labeled target data, our next research question is:

- How do the results of the self-training strategy with hard-labeled target data compare with that of the EM strategy with soft-labeled target data?

Finally, we want to see how the results of the domain adaptation approach compare with the results of ideal supervised learning classifiers trained from target labeled data, with the assumption that time and effort would be spent to manually label the available unlabeled target data. Thus, our last research question is:

- How close are the results of the domain adaptation classifiers to the results of supervised classifiers learned from a large amount of target labeled data?

To have a comprehensive set of results, we select a variety of pairs of disasters when performing experiments, as shown in Table 3. The pairs are selected based on the timeline of the disasters, and also based on the nature (or type) of the disasters, as we want to include pair of disaster of the same type and of different types. Since Hurricane Sandy is the earliest disaster among the six disasters in the CrisisLexT6 dataset, it is never used as target. However, it is used as source in the first five pairs (P1-P5), while the other five disasters are used as target, respectively. The goal here is to experiment with a fixed source but varying targets with increasing distance in time from the source. The experiments with these pairs can tell us if the distance in time from the source disaster to the target disaster matters. Furthermore, we experiment with pairs of disasters of different types (P6-P9), and also with pairs of disasters of the same type (P10-P11). Intuitively, regardless of the distance in time, the similarity in terms of the type of the disaster should generally help.

For each pair of source and target disasters, we use 5-fold cross-validation to select the target data to be used as unlabeled and test data. Specifically, the target data is split into five folds; at each rotation, one fold is selected as target test (*TT*), and three folds as target unlabeled data (*tTU*), used by the domain adaptation approach with self-training/expectation-maximization, together with source labeled data (*tSL*). The fifth fold is reserved as target labeled data to be used in future work.

Table 3. Pairs of disasters used in the experiments. The first group of pairs (P1-P5) has Hurricane Sandy as source and the other disasters as target, respectively. The second group of pairs (P6-P9) consists of pairs with different types of disasters as source and target, respectively. The last group of pairs (P10-P11) consists of pairs with the same type of disaster for both source and target.

Pair	**Source Disaster Set**	**Target Disaster Set**
P1	2012_Sandy_Hurricane	2013_Queensland_Floods
P2	2012_Sandy_Hurricane	2013_Boston_Bombings
P3	2012_Sandy_Hurricane	2013_West_Texas_Explosion
P4	2012_Sandy_Hurricane	2013_Oklahoma_Tornado
P5	2012_Sandy_Hurricane	2013_Alberta_Floods
P6	2013_Queensland_Floods	2013_Boston_Bombings
P7	2013_Queensland_Floods	2013_Oklahoma_Tornado
P8	2013_Boston_Bombings	2013_Oklahoma_Tornado
P9	2013_Boston_Bombings	2013_Alberta_Flood
P10	2013_Queensland_Floods	2013_Alberta_Floods
P11	2013_Boston_Bombings	2013_West_Texas_Explosion

We report the results using accuracy and area under the receiver operating characteristic curve (*auROC*) averaged over the five target test folds. Both accuracy and *auROC* are metrics commonly used in machine learning, and capture different qualities of a classifier. The accuracy measures the percentage of correctly labeled instances out of the total number of instances. The *auROC* measures the ability of the classifier to rank instances based on the probability, $P(c|d)$, that they belong to a class (positive or negative), without effectively assigning instances to classes. Given a ranking, the ROC plots the true positive rate as a function of the false positive rate, obtained at different cut-points in the ranking. As opposed to that, the accuracy is obtained based on one single cut-point (most commonly 0.5).

For each pair of source-target disasters, we perform four groups of experiments as described below, one for each of our research questions stated above.

Supervised learning from source labeled data only. In this group of experiments, we use source labeled data as the training set, and learn supervised Naïve Bayes classifiers. We then use the resulting classifiers to classify target test data. Thus, the classifiers learned in this group of experiments can serve as baselines (intuitively, lower bounds) for the domain adaptation classifiers. We denote the supervised Naïve Bayes classifiers by *NB-S*. The training data for this classifier, training source labeled data, is denoted by *tSL*. Given that other supervised classifiers have been used successfully in prior work, we also compare the results of the supervised Naïve Bayes classifiers with the results of supervised random forest (RF), logistic regression (LR) and support vector machine (SVM) classifiers. One advantage of the Naïve Bayes classifier over other classifiers, and the reason our domain adaptation approach uses it as base classifier, is that the Naïve Bayes algorithm does not have any parameters that require tuning. We used an open-source machine learning library, called WEKA (Hall *el at.*, 2009), to learn supervised classifiers. We used default parameters for the RF, LR, SVM algorithms.

Domain adaptation with Self-Training and Expectation-Maximization, respectively. There are two groups of domain adaptation experiments. One is domain adaptation with self-training and hard-labeled target data, the other is domain adaptation with expectation-maximization and soft-labeled target data. In both groups, we use source labeled data and target unlabeled data to train a domain adaptation classifier for the target, and subsequently test the classifier on the target test data. We use the notation *NB-EM* for domain adaptation with the EM strategy, and *NB-ST* for domain adaptation with the ST strategy. The training data for this classifier is denoted by *tSL+tTU* to suggest that it is consists of source labeled data and target unlabeled data.

Supervised learning from target labeled data. In this group of experiments, we use the target unlabeled dataset (*tTU*) that is used in the domain adaptation setting and assume that the labels of the instances in this dataset are provided. We learn Naïve Bayes classifiers from the target labeled data and test them on target test data. Intuitively, if labeled training data from a target disaster is available, we should be able to learn accurate classifiers for that disaster. Therefore, the results of the supervised classifiers learned from the assumed target labeled data can be seen as upper bounds for all the results of the other classifiers. We denote this supervised Naïve Bayes classifiers by *NB-T**.

Parameter tuning. The domain adaptation approach has two parameters that need to be tuned: the parameter δ that controls how fast we shift the weight from source to target in both ST and EM strategies; and the parameter *k* that controls how many instances to hard-label at each iteration of the ST strategy. To avoid over-fitting, we tune parameters during a validation step. For the validation, we select one of three target unlabeled (*tTU*) folds as validation data (*TV*), and use the other two folds of *tTU* as target unlabeled data (*tUV*). We use *tSL+tUV* for training and test on *TV* to select the best values for the parameters. After tuning, the whole *tTU* is used for self-training as well as for EM. The performance metrics are estimated using the target test set *TT*. The following values are considered for the parameter δ: {0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9}. In addition, we hard-label *k*=1, 5, or 10 instances of on-topic and off-topic, respectively.

6 Experimental Results and Discussion

Tables 4a and 4b show the results of the supervised classifiers learned from source data only in terms of accuracy and *auROC* (averaged over 5-folds), respectively. We compare the following classifiers: Naïve Bayes (NB-S), Support Vector Machines (SVM-S), Random Forests (RF-S) and Logistic Regression (LR-S). As mentioned earlier, we used the Weka implementations (Hall *el at.*, 2009), with default parameters. Furthermore, we used the LibLinear variant of SVM (Fan *et al.*, 2008), as opposed to LibSVM variant (Chang and Lin, 2011), as the results were consistently better for the LibLinear variant. We performed paired t-tests to identify classifiers that are statistically significantly better than their counterparts (using p<0.05). The best values within the four rows of a pair are shown in bold.

Table 4a. Accuracy results (averaged over 5-folds) for the eleven pairs of disasters (P1-P11) and four supervised classifiers trained from labeled source data only: supervised Naive Bayes using source only as training set (*NB-S*); supervised support vector machines (*SVM-S*); supervised random forest (*RF-S*); supervised logistic regression (*LR-S*). *SVM-S*, *RF-S* and *LR-S* are trained with default Weka parameters. The best values (according to a t-test with p<0.05) obtained with any of four classifiers for each pair are shown in bold font.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
NB-S	**76.84**	68.66	**77.21**	**80.78**	**71.06**	**74.97**	**84.13**	**84.35**	**73.81**	**78.87**	**94.77**
SVM-S	73.81	55.23	63.43	77.15	68.86	65.76	**83.97**	81.45	71.17	76.69	84.29
RF-S	70.11	**73.33**	**77.16**	**79.78**	65.85	71.65	81.56	82.45	**73.95**	74.49	92.15
LR-S	71.85	56.41	64.86	72.60	67.09	58.00	79.01	79.78	66.42	72.06	87.85

Table 4b. Weighted *auROC* results (averaged over 5-folds) for the eleven pairs of disasters (P1-P11) and four supervised classifiers trained from labeled source data only: supervised Naive Bayes using source only as training set (*NB-S*); supervised support vector machines (*SVM-S*); supervised random forest (*RF-S*); supervised logistic regression (*LR-S*). *SVM-S*, *RF-S* and *LR-S* are trained with default Weka parameters. The best values (according to a t-test with p<0.05) obtained with any of four classifiers for each pair are shown in bold font.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
NB-S	**0.911**	0.753	0.853	**0.865**	**0.830**	0.820	**0.880**	**0.905**	**0.806**	**0.860**	**0.983**
SVM-S	0.763	0.555	0.626	0.743	0.707	0.661	0.824	0.810	0.705	0.733	0.835
RF-S	0.885	**0.825**	**0.873**	**0.860**	0.818	**0.833**	**0.899**	0.891	0.818	**0.860**	0.977
LR-S	0.847	0.525	0.598	0.612	0.766	0.472	0.775	0.817	0.628	0.714	0.919

As can be seen from Tables 4a and 4b, the Naïve Bayes classifier has the overall best performance in terms of both accuracy and *auROC* metrics, when compared with other supervised classifiers trained with default parameters. Furthermore, the Naïve Bayes classifier has the advantage that it does not require any parameter tuning. Given these reasons, we build our domain adaptation classifiers based on Naïve Bayes, and compare the domain adaptation classifiers only with supervised Naïve Bayes classifiers in what follows.

The results of the domain adaptation classifiers by comparison with supervised Naïve Bayes classifiers are shown in Tables 5a and 5b in terms of accuracy and *auROC*, respectively.

Table 5a. Accuracy results (averaged over 5-folds) for the eleven pairs of disasters (P1-P11) and four approaches: supervised Naive Bayes using source only as training (*NB-S*); domain adaptation with EM (*NB-EM*); domain adaptation with ST (*NB-ST*); supervised Naive Bayes using target unlabeled data as labeled training data (*NB-T**). The results of the *NB-T** classifiers are underlined to suggest that they can be seen as an upper bound for the other classifiers for a pair. The best values (according to a t-test with p<0.05) obtained with any of the first three approaches for each pair are shown in bold font.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
NB-S	76.84	68.66	77.21	80.78	71.06	74.97	84.13	84.35	73.81	78.87	94.77
NB-EM	78.96	80.88	**94.66**	**87.58**	76.87	76.69	**86.63**	**87.44**	**82.47**	82.43	**95.79**
NB-ST	**82.40**	**84.06**	90.82	**87.76**	**82.57**	**81.86**	85.48	**86.91**	**83.96**	**86.01**	94.82
NB-T*	*93.42*	*89.20*	*95.90*	*90.45*	*92.63*	*89.20*	*90.45*	*90.45*	*92.63*	*92.63*	*95.90*

Table 5b. Weighted *auROC* (averaged over 5-folds) for the eleven pairs of disasters (P1-P11) and four approaches: supervised Naive Bayes using source only for training (*NB-S*); domain adaptation with EM (*NB-EM*); domain adaptation with ST (*NB-ST*); supervised Naive Bayes using the target unlabeled data as training labeled data (*NB-T**). The results of the *NB-T** classifiers are underlined to suggest that they can be seen as an upper bound for the other classifiers for a pair. The best values (according to a t-test with p<0.05) obtained with any of the first three approaches for each pair are shown in bold font. If the best values are equivalent to the value obtained with the corresponding supervised classifier NB-T*, they are also underlined. Furthermore, values that are better than the value obtained with NB-T* are also marked with a star (*).

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
NB-S	0.911	0.753	0.853	0.865	0.830	0.820	0.880	0.905	0.806	0.860	0.983
NB-EM	**0.973***	0.929	**0.984**	**0.938**	0.953	0.832	**0.925**	**0.942**	0.898	0.882	**0.989**
NB-ST	**0.974***	**0.941**	**0.987**	**0.951**	**0.972**	**0.890**	**0.924**	**0.944**	**0.950**	**0.922**	0.984
*NB-T**	*0.969*	*0.954*	*0.989*	*0.961*	*0.971*	*0.954*	*0.961*	*0.961*	*0.971*	*0.971*	*0.989*

Table 5a shows the 5-fold average accuracy for each classifier and each pair. Table 5b shows the 5-fold average weighted *auROC* for each classifier and each pair. The first row in each of the result tables corresponds to the supervised Naïve Bayes classifiers learned from source only (*NB-S*), the next two rows correspond to the domain adaptation approaches with Expectation-Maximization (*NB-EM*) and Self-Training (*NB-ST*), respectively. The last row in each table *NB-T** corresponds to an ideal classifier learned from target labeled data. The results of this classifier, underlined, can be seen as an upper bound for the results that can be achieved with domain adaptation which has access to only unlabeled data from the target domain, in addition to labeled data from a prior source domain. For a more visual comparison, Figures 1a and 1b show the results of the first three rows of Tables 5a and 5b, respectively, using bar charts.

As for the supervised classifiers, we performed paired t-tests to identify classifiers that are statistically significantly better than their counterparts (using $p<0.05$). The best values within the first three rows of a pair are shown in bold. Furthermore, if a result is equivalent to the result of the ideal *NB-T** classifier, it is indicated with underscore, and underscore with star (*) means that the corresponding domain adaptation result is better than the result of the *NB-T** classifier.

Based on the results in Tables 4 and 5, we answer our research questions below.

How do the supervised classifiers learned only from source labeled data perform on target data? As our results in Tables 4a and 4b show, labeled data from a prior source disaster can be very useful for learning classifier for different target disasters. When using only source labeled data to learn Naïve Bayes classifier (the approach *NB-S*), the *auROC* values are greater than 0.8 or 0.9 for most pairs, with the exception of pair P2, for which the *auROC* value is around 0.75. Similarly, the accuracy for most pairs is over 70% or 80% except for pair P2 as well. The accuracy and *auROC* values are especially high when considering disasters of the same type (e.g., pair P11: Boston_Bombings -> West_Texas_Explosion), and relatively smaller for more different disasters (e.g., pairs P6: Quessland_Floods->Boston_Bombings, and P9: Boston_Bombings->Alberta_Floods). Furthermore, it is worth noting that, while both pairs P5 and P10 have Alberta Floods as target, the results for P10, which has Queensland Floods as source (another flood) are better than the results for P5, which has Sandy Hurricane as source. Similar behavior is observed for pairs P3 and P11, which both have West Texas Explosion as target: Boston Bombings as source in P11 gives better results than Sandy Hurricane in P3. Together, these results show that supervised learning based on source can be used to learn classifiers for a target if the source and target disasters are similar. This conclusion is consistent with other prior studies (Verma *et al.* 2011, Imran *et al*. 2013b, Li *et al.* 2015).

A more interesting observation is that for the pair P11, the supervised Naïve Bayes classifier is highly accurate, with accuracy close to 95% and *auROC* close to 1.0. By examining sample tweets from the two disasters, we find that they share more common features than other pairs of disasters in our experiments. Reasons for the common features include the fact that the West Texas Explosion happened shortly after the Boston Bombings, both disasters happened in US, and they were man-made. Thus, people who tweeted about West Texas Explosion often mentioned Boston Bombings as well. However, this is not the case for the pair P10 (Queensland_Floods->Alberta_Floods), where both the source and the target disasters are floods (natural disasters), with different geo-locations, but people don't talk about Alberta_Floods in relation to Queensland_Floods.

Another interesting observation can be made for pairs P2 and P6 that have Boston Bombings as target, but Hurricane Sandy (P2) versus Queensland Floods (P6) as sources, respectively. As Hurricane Sandy mostly affected the east coast of the US, one might expect that the classifier for pair P2 may give better accuracy than the classifier for pair P6.

However, this is not the case as can be seen in Table 4a, which suggests that domain similarity or closeness of disasters can be more sensitive to the occurring times of the disasters rather than geo-locations, or other facts about the disaster types. More datasets and experiments are needed to get a firmer conclusion in this respect.

Figure 1a. Accuracy results (averaged over 5-folds) for the eleven pairs of disasters (P1-P11) and three approaches: supervised Naive Bayes using source only as training data (*NB-S*); domain adaptation with Expectation-Maximization (*NB-EM*); domain adaptation with Self-Training (*NB-ST*).

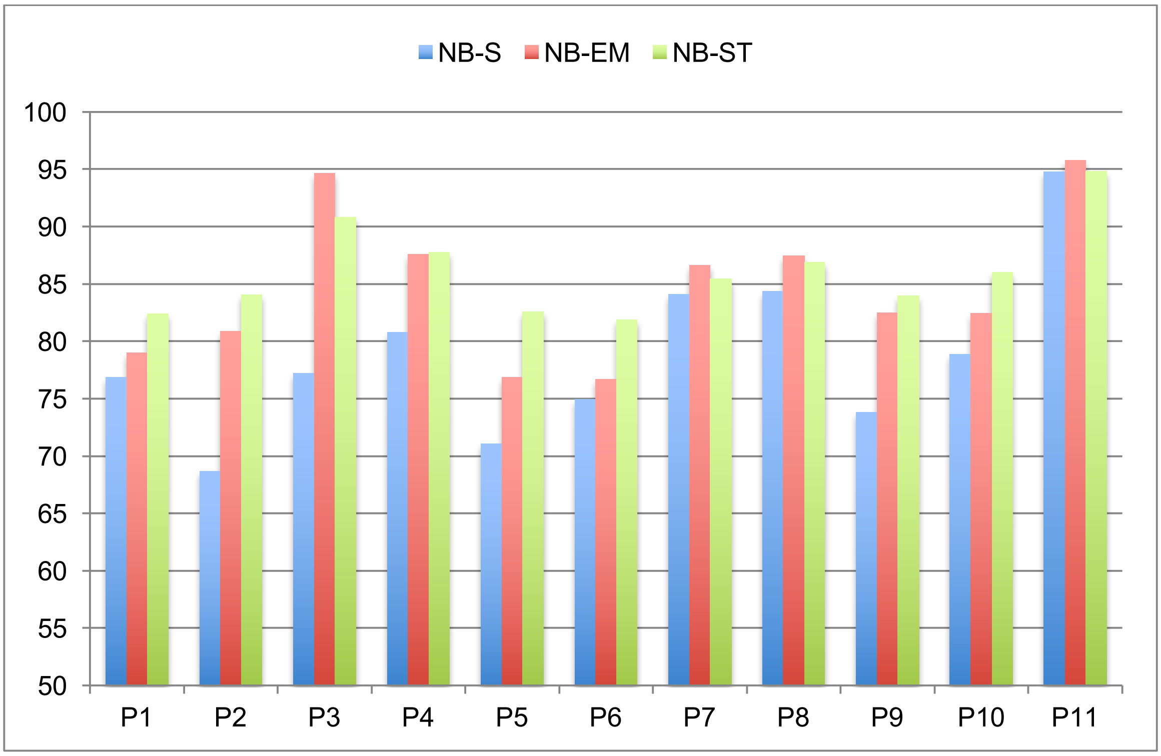
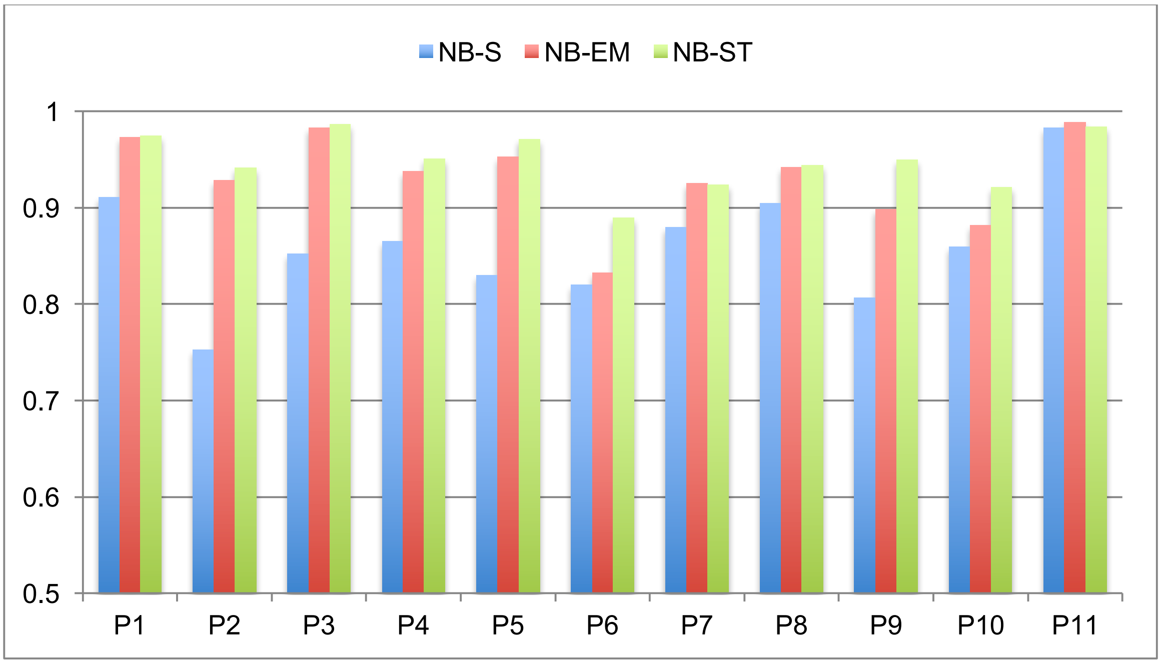


Figure 1b. Weighted *auROC* (averaged over 5-folds) results for the eleven pairs of disasters (P1-P11) and three approaches: supervised Naive Bayes using source only as training data (*NB-S*); domain adaptation with Expectation-Maximization (*NB-EM*); domain adaptation with Self-Training (*NB-ST*).



How do the results of the domain adaptation classifiers that use both source labeled data and target unlabeled data compared with the results of the supervised classifiers that use only source data, when used to classify target data? As can be seen from Tables 5a and 5b, domain adaptation approaches that make use of target unlabeled data definitely help to improve the results of the classifiers learned from source data only. By comparing the results of the supervised *NB-S* with the results of the domain adaptation approaches *NB-EM* and *NB-ST*, we can see that for the first 10 pairs of disasters considered (P1-P10), domain adaptation classifiers with either EM or ST are better than the corresponding supervised classifiers. For some pairs, the improvement is very big; for example, for pairs P2, P3, P5 and P9, the accuracy in Table 5a has improved by more than 10% when using the domain adaptation approach *NB-ST* as compared to the supervised learning algorithm with source only. For pair P11, domain adaptation *NB-EM* with soft-labels still improves the accuracy, whereas domain adaptation *NB-ST* with hard-labels doesn't help much. The reasons for this may lie in the fact that the source itself is close to that target, and the instances added with self-training are not very different from the source instances. Still the domain adaptation with EM give results that are better than the results of the supervised classifier, according to the *t*-test (at p<0.05).

How do the results of the self-training strategy with hard-labeled data compare with those of the EM strategy with soft-labeled target data? When comparing the *NB-ST* approach (with hard-labels) with the *NB-EM* approach (with soft-labels), we can see that in general *NB-ST* performs better than *NB-EM*. More specifically, for 8 out of 11 pairs, *NB-ST* is either equivalent (3 pairs) or better (5 pairs) than *NB-EM* in terms of accuracy, and for all pairs except P11, either equivalent or better in terms of *auROC*. For pair P3, although the accuracy in Table 5a is higher for *NB-EM*, the weighted *auROC* in Table 5b is equivalent to *NB-ST*. *NB-EM* with soft-labels is statistically better than *NB-ST* with hard-labels only on pairs P3, P7 and P11. EM which is using all target unlabeled data at each iteration may provide more information than ST as the sources in pair P3 and P7 are not only of different types as compared to the target, but also far in time, and thus the original classifier learned from them is not reliable enough to accurately label a small number of instances for the *NB-ST* approach. However, overall, we can confidently say that ST performs better than EM for our classification task.

How close are the results of the domain adaptation classifiers to the results of supervised classifiers learned from a large amount of target labeled data? Finally, to answer our last research question, from Tables 5a and 5b, we can see that domain adaptation approaches can achieve results very close to the upper bound in several cases but not always. By comparing the accuracy results of EM/ST in Table 5a with the accuracy results obtained with the ideal *NB-T** approach (used as an upper bound), we can see that the domain adaptation algorithms get close to the upper bound in a few cases, for example, for pairs P3, P4, P5 and especially pair P11. However, in general, there is still significant room for improving the accuracy results of the domain adaptation classifiers. By comparing the *auROC* results in Table 5b, we can see that the results of the domain adaptation classifiers are, in general, closer to the upper bounds except for pairs P6 and P7. Furthermore, in some cases the results are even better than the upper bound, for example for pair P1. This can be explained by the fact that the source itself provides accurate results (*auROC* value higher than 0.9), which makes it possible to accurately label the originally unlabeled target data. Thus, the accurately labeled target data together with the source data produce classifiers that are better than those learned from labeled target data alone (which could be noisy).

7 Conclusion

In this article, we studied an automated solution for sifting through increasingly overwhelming amounts of data contributed directly by communities affected by a disaster. Our solution is based on a domain adaptation approach, adapted from (Li *et al*., 2015), which makes use of a self-training iterative strategy to incorporate labeled data from a source disaster and unlabeled data from an emerging target disaster into a classifier for the target disaster.

We used a relatively large dataset, crisis tweets dataset from Olteanu *et al*. (2014), to evaluate our proposed domain adaptation classifiers based on Naïve Bayes and self-training with hard labels (*NB-ST*), and to compare them with supervised classifiers learned only from source (*NB-S*) and with domain adaption classifiers based on expectation-maximization (*NB-EM*). The results of our experiments showed that using source data only with supervised learning can help when the source and target disasters are similar. However, the domain adaptation approaches are always better than the supervised learning with source data only. Between the *NB-ST* and *NB-EM* approaches, generally the *NB-ST* approach is better. As last, our experimental results showed that the domain adaptation approaches can give results comparable, and in some cases better, than an ideal supervised classifier that would have (noisy) labeled target data available. However, in general, there is still room for improving the results of domain adaptation classifiers as compared to the ideal supervised classifier that would have access to labeled target data.

By helping analyze millions of tweets automatically, our proposed approach has the potential to impact the way in which response organizations operate, particularly, by identifying more accurate and timely information than it is possible with traditional information gathering methods, and in turn providing better support to those who need them, and even, saving more lives.

As part of the future work, we would like to study how the NB-ST algorithm performs for multi-class tasks, as opposed to binary-class tasks, for example classification of tweets with respect to situational awareness categories such as: donation, casualty, damage, etc. We would also like to investigate multi-source domain adaptation algorithms and apply them in disaster management as well. In general, the labeled data from a particular disaster is relatively small, especially for disasters that have data labeled with very specific information, such as infrastructure damage, donation, etc. Multi-source domain adaptation, which lets us make use of multiple sources of data, can potentially help to address this problem.

Acknowledgements

The computing for this project was performed on the Beocat Research Cluster at Kansas State University. We thank the National Science Foundation for support from grant CNS-1429316, which funded the cluster. We also thank the National Science Foundation for support from the grants IIS-1526542 and CMMI-1541155 to Cornelia Caragea. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either express or implied, of the National Science Foundation. We also wish to thank our anonymous reviewers for their constructive comments.

References

Ashktorab, Z., Brown, C., Nandi, M. & Culotta, A. (2014) Tweedr: Mining Twitter to Inform Disaster Response. In *Proceedings of 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2014)* (pp.354-358), University Park, PA, USA.

BBC Trending(2015). BBC News. http://www.bbc.com/news/blogs-trending-34836214.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2016). An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment Analysis and Ontology Engineering* (pp. 313-340). Springer International Publishing.

Blum, A., & Mitchell, T. (1998). Combining Labeled and Unlabeled Data with Co-training. In *Proceedings of the eleventh annual conference on Computational learning theory* (pp. 92-100). ACM.

Caragea, C., Silvescu, A., & Tapia, A. H. (2016). Identifying informative messages in disaster events using convolutional neural networks. In *Proceedings of 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2016)*. Rio de Janeiro, Brazil, May 2016.

Caragea, C., Squicciarini, A., Stehle, S., Neppalli, K., & Tapia, A. (2014) Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy. In *Proceedings of 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2014)* (pp. 642-651), University Park, PA, USA.

Castillo, C. (2016). Big Crisis Data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press.

Chang,C. & Lin,C.(2011) LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27, 2011.

Dai, W., Xue G., Yang, Q., & Yu, Y. (2007) Transferring Naïve Bayes Classifiers for Text Classification. In *Proceedings of the AAAI 2007 Conference on Artificial Intelligence* (pp. 540-545), Vancouver, British Columbia, Canada.

Dugdale, J., Van de Walle, B., & Koeppinghoff, C. (2012). Social media and SMS in the haiti earthquake. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 713-714). ACM,.

Fan, R. E., Chang, K. W., Hsieh, C. J., Wang, X. R., & Lin, C. J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of machine learning research*, 9(Aug), 1871-1874.

Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, *26*(3), 10-14.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: an Update. *ACM SIGKDD explorations newsletter*, 11(1), 10-18.

Harvard Humanitarian Initiative (2011). Disaster relief 2.0: The future of information sharing in humanitarian emergencies. Washington, DC and Berkshire, UK: UN Foundation & Vodafone Foundation Technology Partnership.

Herndon, N. & Caragea, D. (2014) Empirical Study of Domain Adaptation with Naïve Bayes on the Task of Splice Site Prediction. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS)* (pp. 57–67)*., ESEO, Angers, Loire Valley, France.*

Herndon, N. & Caragea, D. (2015). An Evaluation of Self-training Styles for Domain Adaptation on the Task of Splice Site Prediction. In *Proceedings of the 2015 International Symposium on Network Enabled Health Informatics, Biomedicine and Bioinformatics (HI-BI-BI 2015)* (pp. 1042-1047), Paris, France.

Homeland Security (2014). Using Social Media for Enhanced Situation Awareness and Decision Support. *Virtual Social Media Working Group and DHS First Responders Group*.

Hughes, A., Denis, L. S., Palen, L., & Anderson, K. (2014). Online Public Communications by Police & Fire Services during the 2012 Hurricane Sandy. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'14)* (pp.1505-1514), Toronto, Canada.

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.

Imran, M., Elbassuoni,S., Castillo, C., Diaz, F. & Meier, P. (2013) Practical Extraction of Disaster-Relevant Information from Social Media. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 1021-1024), Rio de Janeiro, Brazil.

Imran, M., Elbassuoni,S., Castillo, C., Diaz, F. & Meier, P. (2013) Extracting Information Nuggets from Disaster-Related Messages in Social Media. In *Proceedings of 10th the International Conference on Information Systems for Crisis Response and Management (ISCRAM 2013)* (pp.791-800), Baden-Baden, Germany.

Imran, M., Mitra, P., & Srivastava, J. (2016). Cross-Language Domain Adaptation for Classifying Crisis-Related Short Messages. In *Proceedings of 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2016)*. Rio de Janeiro, Brazil, May 2016.

Jiang, J. (2008). A literature survey on domain adaptation of statistical classifiers. URL: http://sifaka. cs. uiuc. edu/jiang4/domainadaptation/survey.

Kaufhold, M. A., & Reuter, C. (2016). The Self-Organization of Digital Volunteers across Social Media: The Case of the 2013 European Floods in Germany. J*ournal of Homeland Security and Emergency Management*, 13(1), 137-166.

Kumar, S., Hu, X., & Liu, H. (2014). A Behavior Analytics Approach to Identifying Tweets from Crisis Regions. In *Proceedings of the 25th ACM conference on Hypertext and social media (HT '14)* (pp. 255-260) ACM, New York, NY, USA.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th international conference on World wide web (WWW '10)* (pp. 591-600), New York, NY, USA.

Landwehr, P. M. & Carley, K. M. (2014). Social Media in Disaster Relief: Usage Patterns, Data Mining Tools, and Current Research Directions. In *Data Mining and Knowledge Discovery for Big Data Studies*, Wesley, W.C. (ed.) (pp. 225-257). Springer Heidelberg New York Dordrecht London.

Lewis, D. D. 1992. Representation and Learning in Information Retrieval. Ph.D. Dissertation, Amherst, MA, USA.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A., & Tapia, A. (2015). Twitter Mining for Disaster Response: A Domain Adaptation Approach. In *Proceedings of the 12th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2015)*, Kristiansand, Norway.

Manning, C.D., Raghavan, P.& Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Meier, P (2013). Crisis Maps: Harnessing the Power of Big Data to Deliver Humanitarian Assistance. *Forbes Magazine*, May 2, 2013. Retrieved from https://www.forbes.com/sites/skollworldforum/2013/05/02/crisis-maps-harnessing-the-power-of-big-data-to-deliver-humanitarian-assistance/#73b91c24115c.

Meier, P. (2015). Digital humanitarians: how big data is changing the face of humanitarian response. Crc Press.

Mendoza, M., Poblete, B., & Castillo, C. (2010). Twitter under Crisis: Can We Trust What We RT? In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*(pp. 71-79), New York, NY, USA.

Munro, R. (2011). Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL '11)*(pp. 68-77). Association for Computational Linguistics, Stroudsburg, PA, USA.

Murdock, S (2015). Parisians Can Use A Twitter Hashtag To Seek Shelter During Terrorist Attacks. Huffington Post, Retrieved November 15, 2015, from http://www.huffingtonpost.com/entry/paris-france-terrorist-attack-trending-twitter_5646676ee4b0603773491b7a.

Nigam, K., & Ghani, R. (2000). Analyzing the Effectiveness and Applicability of Co-training. In *Proceedings of the ninth international conference on Information and knowledge management* (pp. 86-93). ACM.

Olteanu, A., Castillo, C., Diaz, F. & Vieweg, S.(2014). CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM'14)*. AAAI Press, Ann Arbor, MI, USA.

Palen, L., & Anderson, K. M. (2016). Crisis informatics—New data for extraordinary times. *Science*, 353(6296), 224-225.

Palen, L., & Vieweg, S. (2008). The Emergence of Online Wide Scale Interaction in Unexpected Events: Assistance, Alliance & Retreat. In *Proceedings of the ACM 2008 Conference on Computer supported cooperative work (CSCW2008)* (pp. 117-126), ACM Press.

Palen, L., Vieweg, S., & Anderson, K. M. (2010). Supporting “everyday analysts” in safety- and time-critical situations. *The Information Society*, 27(1), 52-62.

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) (pp.467).

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on* *knowledge and data engineering*, 22(10), 1345-1359.

Peddinti, V. & Chintalapoodi, P.(2011) Domain Adaptation in Sentiment Analysis of Twitter. *Analyzing Microtext, Papers from the 2011 (AAAI) Workshop,* San Francisco,California, USA.

Purohit, H., Castillo, C., Diaz, F., Sheth, A. & Meier,P. (2013). Emergency-relief Coordination on Social Media: Automatically Matching Resource Requests and Offers. *First Monday*, 19(1), 2013.

Qadir, J., Ali, A., ur Rasool, R., Zwitter, A., Sathiaseelan, A., & Crowcroft, J. (2016). Crisis analytics: big data-driven crisis response. *Journal of International Humanitarian Action*, 1(1), 12.

Reuter, C., Ludwig, T., Friberg, T., Pratzler-Wanczura, S., & Gizikis, A. (2015). Social Media and Emergency Services?: Interview Study on Current and Potential Use in 7 European Countries. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 7(2), 36-58.

Rogers, K (2015). Twitter Cats to the Rescue in Brussels Lockdown. The New York Times, http://www.nytimes.com/2015/11/24/world/europe/twitter-cats-to-the-rescue-in-brussels-lockdown.html?_r=0.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010) Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)* (pp.851–860), Raleigh, NC, USA.

Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010) Chatter on the Red: What Hazards Threat Reveals About the Social Life of Microblogged Information. In *Proceedings of the ACM 2008 Conference on Computer supported cooperative work (CSCW 2010)* (pp.241–250), New York, NY, USA.

Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009) Adapting Naïve Bayes to Domain Adaptation for Sentiment Analysis. In *Proceedings of the Advances in Information Retrieval, 31th European Conference on (IR)Research, (ECIR 2009)* (pp.337-349), Toulouse, France.

Tapia, A. H., & Moore, K. (2014). Good enough is good enough: Overcoming disaster response organizations' slow social media data adoption. *Computer Supported Cooperative Work (CSCW)*, 23(4-6), 483-512.

Tapia, A. H., Bajpai, K., Jansen, B. J., Yen, J., & Giles, L. (2011). Seeking the Trustworthy Tweet: Can Microblogged Data Fit the Information Needs of Disaster Response and Humanitarian Relief Organizations. In *Proceedings of 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011)* (pp. 1-10), Lisbon, Portugal.

Tapia, A. H., Moore, K. A., & Johnson, N. J. (2013). Beyond the Trustworthy Tweet: A Deeper Understanding of Microblogged Data Use by Disaster Response and Humanitarian Relief Organizations. In *Proceedings of 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2013)* (pp. 770-778). Baden-Baden.

Terpstra, T. (2012) Towards a realtime Twitter analysis during crises for operational crisis management. In *Proceedings of 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2012)* (pp.1–9), Vancouver, Canada.

Terpstra, T., De Vries, A., Stronkman, R., & Paradies, G. L. (2012). *Towards a realtime Twitter analysis during crises for operational crisis management* (pp. 1-9). Simon Fraser University.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M. & Anderson, K. M. (2011). Natural Language Processing to the Rescue? Extracting" Situational

Awareness" Tweets During Mass Emergency. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM2011),* Barcelona, Catalonia, Spain.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010) Microblogging during Two Natural Hazards Events. In *Proceedings of the the Conference on Human Factors in Computing Systems (CHI 2010)* (pp.1079–1088). New York, NY,USA.

Walton, R., Mays, R., & Haselkorn, M. (2011). Defining "fast": Factors affecting the experience of speed in humanitarian logistics. In *Proceedings of the 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2011)* (pp. 1-10).

Watson, H., Finn, R. L., & Wadhwa, K. (2017). Organizational and Societal Impacts of Big Data in Crisis Management. *Journal of Contingencies and Crisis Management*, 25(1), 15–22.

Yarowsky, D. (1995). Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics* (pp. 189-196). Association for Computational Linguistics.

Yin, J., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2012). Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, *27*(6), 52-59.