Twitter Mining for Disaster Response: A Domain Adaptation Approach

Hongmin Li
Nicolais Guevara
Kansas State University
hongminli@ksu.edu
nicolais@phys.ksu.edu

Nic Herndon
Doina Caragea
Kansas State University
nherndon@ksu.edu
dcaragea@ksu.edu

Kishore Neppalli
Cornelia Caragea
University of North Texas
kishoreneppalli@my.unt.edu
cornelia.caragea@unt.edu

Anna Squicciarini
Andrea H. Tapia
Pennsylvania State University
asquicciarini@ist.psu.edu
atapia@ist.psu.edu

ABSTRACT

Microblogging data such as Twitter data contains valuable information that has the potential to help improve the speed, quality, and efficiency of disaster response. Machine learning can help with this by prioritizing the tweets with respect to various classification criteria. However, supervised learning algorithms require labeled data to learn accurate classifiers. Unfortunately, for a new disaster, labeled tweets are not easily available, while they are usually available for previous disasters. Furthermore, unlabeled tweets from the current disaster are accumulating fast. We study the usefulness of labeled data from a prior *source* disaster, together with unlabeled data from the current *target* disaster to learn *domain adaptation* classifiers for the target. Experimental results suggest that, for some tasks, source data itself can be useful for classifying target data. However, for tasks specific to a particular disaster, domain adaptation approaches that use target unlabeled data in addition to source labeled data are superior.

Keywords

Disaster response, tweet classification, domain adaptation.

BACKGROUND AND MOTIVATION

Much has been written concerning the value of using messaging and microblogged data from crowds of non-professional participants during disasters. Often referred to as microblogging, the practice of average citizens reporting on activities “on-the-ground” during a disaster is seen as increasingly valuable (Homeland Security, 2014; Terpstra, 2012). According to Vieweg *et al*. (2010), microblogging is seen to have intrinsic value across responder organizations and victims because of its growing ubiquity, communications rapidity, and cross-platform accessibility.

However, there are numerous challenges when considering the use of microblogged data, including issues of reliability, quantification of performance, and translation of reported observations into a form that can be combined with other information. Disasters by nature are unpredictable and complex events, which means every disaster is unique with regards to geography, culture, infrastructure, technology, etc. A responder must take in a great amount of information, often faulty and incomplete, and must perform decision-making under stressful conditions. Still, researchers are optimistic about the value of information from social media data provided that issues surrounding relevance can be reasonably resolved (Hughes *et al.*, 2014; Starbird *et al*. 2010).

Machine learning, data mining, and natural language processing have made great leaps in extracting, processing and classifying disaster-related social media data (Ashktorab *et al*., 2014; Caragea *et al*., 2014; Imran *et al*., 2013a; Mendoza *et al*., 2010; Purohit *et al*., 2013; Sakaki *et al*., 2010). For example, Ashktorab *et al*. (2014) used a combination of classification, clustering, and extraction methods to extract actionable information for disaster responders. Caragea *et al*. (2014) performed sentiment classification of user posts in Twitter during Hurricane Sandy and visualized these sentiments on a geographical map centered on the hurricane location. Imran *et al*. (2013) first used a classifier to identify informative tweets in a dataset collected during the Joplin 2011 tornado, and subsequently classified the informative tweets into more specific types, such as *casualties and damage*, *donations*, etc. Finally, they extracted information nuggets such as *location*, *time*, etc. for different types of tweets.

While machine learning can be used to mine disaster data and help the response teams and victims, it relies on labeled data to learn accurate classifiers. However, when a disaster happens, no labeled data is available for that particular disaster, making it difficult to use machine learning to train models. To address this limitation, we propose to use domain adaptation approaches to learn classifiers either from labeled data from a previous source disaster (for tasks that are similar across different disasters), or from labeled source data together with unlabeled target data (for tasks that are more different across disasters).

Domain adaptation has been extensively used in text classification and sentiment analysis. For example, Dai *et al*. (2007) proposed a domain adaptation algorithm based on Naïve Bayes and Expectation-Maximization (EM), to classify text documents into several categories.

Tan *et al.* (2009) proposed a weighted version of the multinomial Naïve Bayes classifier combined with EM, for sentiment analysis. In the first step, they train a Naïve Bayes classifier on the source data and label the unlabeled data from the target domain. In subsequent steps, they use the EM algorithm with a weighted combination of the source and target data to train the Naïve Bayes classifier. The EM steps are repeated until convergence, with the weight shifted from source to target domain at each iteration.

Herndon and Caragea (2014) developed an approach similar to the approach in (Tan *et al.*, 2009). They also used a weighted Naïve Bayes classifier, combined with the iterative approach of the EM and *self-training* (where only the most confident predictions from the unlabeled target dataset are fed back to the labeled dataset). The approach in (Herndon and Caragea, 2014) was successfully used for bioinformatics problems.

Peddinti and 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.

In the context of disaster response, Imran *et al*. (2013b) explored domain adaptation for the problem of identifying information nuggets using conditional random fields (CRF). They used data from two disasters, Joplin 2011 tornado (as source) and Hurricane Sandy (as target), and learned supervised classifiers 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.

While the work by Imran *et al*. (2013b) represents an important first step towards using domain adaptation for disaster response problems, the algorithms used in this work are all supervised, and cannot make use of unlabeled data readily available for a target disaster. As opposed to that, we use a domain adaptation classifier, similar to the one in (Herndon and Caragea, 2014), that can make use of target unlabeled data, in addition to source data, and study the usefulness of target unlabeled data for learning accurate classifiers for the target. We compare the domain adaptation approach with a supervised approach similar to the one proposed by Imran *et al*. (2013b), where a supervised Naïve Bayes classifier is learned from source only and tested on target.

Intuitively, the supervised classifiers learned from source might perform well on the target for tasks that are similar across the two disasters. However, for more different tasks, the unlabeled target should improve the classifiers learned from source data only. Experimental results using Hurricane Sandy as source and Boston Marathon bombings as target confirm our intuition.

DOMAIN ADAPTATION APPROACH

Our goal is to label tweets from an emergent target disaster using existing labeled data from a previous source disaster. We assume that no training labeled data is available for the target, and study the usefulness of the training source labeled data (*tSL*) by itself, or together with training target unlabeled data (*tTU*) in learning domain adaptation classifiers for the target disaster.

To study the usefulness of the source labeled data by itself, we learn supervised Naïve Bayes classifiers form source only, and use the resulting classifiers for the target. To combine source labeled data with target unlabeled data, we use a domain adaptation approach based on the Naïve Bayes classifier and the iterative approach of the EM. The algorithm is a variant of the algorithm proposed by Herndon and Caragea (2014), with the following modifications:

- As opposed to the algorithm in (Herndon and Caragea, 2014), which uses source labeled data, a small amount of target labeled data and target unlabeled data, we use only source labeled data and target unlabeled data, under the assumption that for an emergent disaster there is no available labeled data.
- As opposed to the previous algorithm, which uses the target labeled data to identify informative features that can bridge the source and the target, here we use all features from the source domain.

DATA COLLECTION AND LABELING

The data used in our experiments are collected from Twitter, using the Twitter API, during the disastrous Hurricane Sandy and during the Boston Marathon bombings, respectively. We randomly selected 1,700 tweets from the Hurricane Sandy collection, and 1,000 tweets from the Boston Marathon bombings collection, and used the Mechanical Turk (MTurk) to manually label each of the randomly selected tweets with respect to three questions:

Q1) Is the tweet about the disaster in question?

Q2) Does the tweet offer support for the victims of the disaster?

Q3) Does the tweet express any emotion to the victims of the disaster?

As can be seen, the first question is more specific to a particular disaster, while the last two questions can be seen as more similar (in terms of predictive features) between different disasters. Each tweet in our Hurricane Sandy and Boston Marathon bombings collections was annotated by two MTurk workers, who had to select one of the following answers for each of the three questions: “Yes”, “No”, “I do not know”. The final labels used in our experiments were obtained by taking the consensus between the two workers, *i.e*., we used only tweets labeled either “Yes” or “No” by both workers. The rest of the tweets were removed from the datasets.

DATA PREPROCESSING

We use the bag-of-words 0/1 representation to represent tweets as vectors of features/words. Before constructing the vocabulary (*a.k.a.*, set of features), we cleaned the tweets as follows:

1. We removed non-printable, ASCII characters, as they are generally regarded as noise rather than useful information.
2. We converted printable HTML entities into their corresponding ASCII equivalents.
3. We replaced URLs, email addresses, and usernames with a *URL*/*email*/*username* placeholder for each type of entity, respectively, under the assumption that for some questions those features could be predictive (e.g., for

Q2 an email address could be indicative of support).

4. We kept numbers, punctuation signs and hashtags, under the assumption that numbers could be indicative of an address (useful for Q2), while punctuation/emoticons and hashtags could be indicative of emotions (useful for Q3).
5. We removed RT (*i.e.*, retweet), under the assumptions that such features are not informative for our classification tasks.
6. Finally, duplicate tweets and empty tweets (that have no characters left after the cleaning) were removed from the data sets.

	Hurricane Sandy			Boston Marathon bombings		
	Yes	No	Ratio	Yes	No	Ratio
Q1	567	149	3.8:1	399	314	1.3:1
Q2	27	1411	1:52	98	662	1:6.8
Q3	49	1200	1:24.5	140	631	1:4.5

Table 1. Statistics for Hurricane Sandy and Boston Marathon bombings.

Table 1 shows the distribution of the remaining tweets for the three questions, for Hurricane Sandy and Boston Marathon bombings disasters, respectively. These are the tweets used in our experiments. For each question and each disaster, we also show the ratio between the two classes (“Yes” and “No”). As can be seen, the datasets corresponding to the first question are relatively balanced. However, the datasets for Q2 and Q3 are somewhat imbalanced for the Boston Marathon bombings, and highly imbalanced for the Hurricane Sandy disaster. The final set of features is obtained from the remaining tweets. However, we filter out features that appear less than 10 times in a collection.

EXPERIMENTAL RESULTS AND DISCUSSION

Our experimental setup is aimed to address the following questions:

1) How useful is the source disaster labeled data for learning classifiers for the target disaster, using a supervised learning algorithm?
2) How useful is the target disaster unlabeled data, in addition to source disaster labeled data, in a domain adaptation framework?

To address these questions, we perform two experiments. In the first experiment, we use the available source labeled data from a previous disaster to train supervised Naïve Bayes classifiers, and use these classifiers to classify test data from the current target disaster. In the second experiment, we use the domain adaptation algorithm to learn classifiers from source labeled data together with target unlabeled data, and use the classifiers to classify the target test data. We use 5-fold cross-validation to generate training and test datasets. At each iteration, a fold is used as target test (*TT*) data, and the remaining four folds are used as training target unlabeled (*tTU*) data. We use the whole training source labeled (*tSL*) data in all experiments. The results, reported using the area under the ROC curve (auROC), are averaged over the five target test folds. Given that the source data is highly imbalanced, we learn classifiers from the original imbalanced source datasets, and also from balanced source datasets, where the balancing is done using over-sampling (which was a natural choice, given the small number of positive tweets available for source, especially for questions Q2 and Q3). However, we want to note that the evaluation is performed on the original target distribution.

We use the Hurricane Sandy as the source disaster and the Boston Marathon bombings as the target disaster, motivated by the chronological order of the two events. The results of our experiments for the three questions, with and without source balancing are shown in Table 2. As can be seen, the classifiers learned from the source data are better than random classifiers (which would have auROC=0.5), in all experiments, but the results are better when balancing the source labeled data, especially for questions Q1 and Q2. Surprisingly, for question Q3 (referring to the tweet expressing any emotions), balancing the source data gives slightly worse results than not balancing. One possible explanation for this could be related to the quality of the labeled data for this question. In general, it might be easier to label tweets with respect to questions Q1 and Q2, than with respect to question Q3 (emotions). If the source labels are noisy, then oversampling will emphasize some wrong labels, and thus can result in worse classifiers than their imbalanced counterparts.

Question	Training data	Weighted Average auROC	
		without balancing	with balancing
Q1	tSL	0.532	0.670
	tSL+tTU	**0.724**	**0.731**
Q2	tSL	0.594	**0.712**
	tSL+tTU	**0.615**	0.627
Q3	tSL	**0.701**	**0.680**
	tSL+tTU	0.664	0.667

Table 2. Experimental results for three questions, when using the Hurricane Sandy as source and Boston Marathon bombings as target, without balancing the source data set and with balancing (specifically, oversampling), respectively. For each question, we compare the supervised classifier learned from source labeled only (tSL), with the domain adaptation classifier learned from source labeled (tSL) and target unlabeled (tTU).

When comparing the supervised classifiers learned from source data only with domain adaptation classifiers learned from source labeled and target unlabeled, we see that for question Q1, which is about relevance to a specific disaster, adding target unlabeled data using domain adaptation can significantly improve the classifiers learned from source labeled data only (both for balanced and imbalanced source data). However, for questions Q2 and Q3, which are more similar across source and target, the supervised source classifiers are generally better. These results correspond to our intuition that target unlabeled data is more useful for tasks specific to a particular disaster (for example, question Q1), as opposed to tasks similar between disasters (questions Q2 and Q3 probably make use of similar language to offer support/identify emotions, respectively, regardless of the specific disaster).

Also, for questions Q2, the supervised source classifiers are much better as compared to those learned from source only when the source labeled data is balanced (which seems intuitive given the high data imbalanced in the source for these questions). However, when adding unbalanced target data, the differences between balanced and imbalanced experiments are not that significant, as the target unlabeled data can compensate to some extent for the effect of the imbalance. However, in some cases, the target unlabeled data can be noisy and deteriorate the classifiers learned from source only, especially when balancing the source and thus increasing its benefits.

Overall, our experiments suggest that source data from a prior disaster can be used to learn classifiers for a current target disaster, especially for tasks that are similar across disasters. Furthermore, using source labeled data together with target unlabeled data in a domain adaptation framework has the potential to produce better classifiers for tasks that are more specific to a disaster.

CONCLUSIONS AND FUTURE WORK

We used Twitter data about Hurricane Sandy and Boston Marathon bombings to study the applicability of domain adaptation algorithms for mining tweets for disaster response. Specifically, under the assumption that target labeled data is not available, we studied two approaches to this problem. In the first approach we learn supervised Naïve Bayes classifiers from the source labeled data only. In the second approach, we use a domain adaptation Naïve Bayes algorithm to learn classifiers from labeled source data together with unlabeled target data. Preliminary experimental results suggest that for tasks that are more specific to the current target disaster, domain adaptation classifiers that use some data from that specific disaster, albeit unlabeled, seem to be superior to classifier learned from source data only. On the other hand, for tasks that are more similar across disasters, classifiers learned from source only perform better, as the target unlabeled data can act as noise in this case.

More experiments with larger datasets, more tasks (especially disaster-specific tasks), and more classifiers (e.g., SVM, random forests) are needed to come up with more general conclusions. In addition, domain adaptation algorithms that make use of a small amount of target labeled data (in addition to source labeled and target unlabeled) will be studied, with the goal of understanding how important it is to label data from the disaster of interest.

ACNOWLEDGMENTS

This research was supported in part by NSF awards #1353418 and #1353400. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of NSF.

REFERENCES

1. Ashktorab, Z., Brown, C., Nandi, M., Culotta, A. (2014) Tweedr: Mining Twitter to Inform Disaster Response. In: *Proceedings of ISCRAM* 2014, 354-358, University Park, PA.
2. Caragea, C., Squicciarini, A., Stehle, S., Neppalli, K., and Tapia, A. (2014) Mapping Moods: Geo-Mapped Sentiment Analysis During Hurricane Sandy. In: *Proceedings of the ISCRAM 2014*, 642-651, University Park, PA.
3. Dai, W., Xue G., Yang, Q., and Yu, Y. (2007) Transferring Naïve Bayes Classifiers for Text Classification. In: *Proceedings of the AAAI 2007 Conference on Artificial Intelligence*, 540-545.
4. Herndon,N. and Caragea,D. (2014) Empirical Study of Domain Adaptation with Naive Bayes on the Task of Splice Site Prediction. In: *Proceedings of the BIOINFORMATICS 2014*, 57–67.
5. Homeland Security (2014) Using social media for enhanced situation awareness and decision support. *Virtual Social Media Working Group and DHS First Responders Group*.
6. Hughes, A., Denis, L. S., Palen, L., and Anderson, K. (2014) Online Public Communications by Police & Fire Services during the 2012 Hurricane Sandy, In: *Proceedings of the CHI 2014*, 1505-1514, Toronto.
7. Imran, M. , Elbassuoni,S., Castillo, C., Diaz, F. and Meier, P. (2013a) Practical Extraction of Disaster-Relevant Information from Social Media. In: *Proceedings of WWW 2013*, 1021-1024, Rio de Janeiro, Brazil.
8. Imran, M. , Elbassuoni,S., Castillo, C., Diaz, F. and Meier, P. (2013b) Extracting Information Nuggets from Disaster-Related Messages in Social Media. In: *Proceedings of the ISCRAM 2013*, 791-800, Baden-Baden, Germany.
9. Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter Under Crisis: Can we trust what we RT? *New York*, ACM Press, 71–79.
10. Palen, L., Vieweg, S., Liu, S. B., and 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), 467.
11. Peddinti, V. and Chintalapoodi, P.(2011) Domain Adaptation in Sentiment Analysis of Twitter. In: Proceedings of AAAI Workshop on Analyzing Microtext, 2011, San Francisco, CA.
12. Purohit, H. , Castillo, C., Diaz, F., Sheth, A. and Meier,P. (2013). Emergency-relief coordination on social media: Automatically matching resource requests and offers. First Monday, 19(1), 2013.
13. Sakaki, T., Okazaki, M., and Matsuo, Y. (2010) Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. *WWW 2010*, 851–860, ACM, Raleigh, NC.
14. Starbird, K., Palen, L., Hughes, A. L., and Vieweg, S. (2010) Chatter on the Red: What Hazards Threat Reveals About the Social Life of Microblogged Information. In: *Proceedings of the CSCW 2010*, 241–250, New York, NY.
15. Tan, S., Cheng, X., Wang, Y., and Xu, H. (2009) Adapting Naïve Bayes to Domain Adaptation for Sentiment Analysis. In: *Proceedings of the ECIR*

2009, 337-349, Berlin, Heidelberg. Springer-Verlag.

16. Terpstra, T. (2012) Towards a realtime Twitter analysis during crises for operational crisis management. In: *Proceedings of ISCRAM 2012*, 1–9.
17. Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010) Microblogging during two natural hazards events. In: *Proceedings of the CHI 2010*, 1079–1088. New York, NY.