A Hybrid Domain Adaptation Approach for Identifying Crisis-Relevant Tweets

Reza Mazloom, Kansas State University, Manhattan, USA  
Hongmin Li, Kansas State University, Manhattan, USA  
Doina Caragea, Kansas State University, Manhattan, USA  
Cornelia Caragea, University of Illinois at Chicago, Chicago, USA  
Muhammad Imran, Qatar Computing Research Institute, Ar-Rayyan, Qatar

ABSTRACT

Huge amounts of data generated on social media during emergency situations is regarded as a trove of critical information. The use of supervised machine learning techniques in the early stages of a crisis is challenged by the lack of labeled data for that event. Furthermore, supervised models trained on labeled data from a prior crisis may not produce accurate results, due to inherent crisis variations. To address these challenges, the authors propose a hybrid feature-instance-parameter adaptation approach based on matrix factorization, k-nearest neighbors, and self-training. The proposed feature-instance adaptation selects a subset of the source crisis data that is representative for the target crisis data. The selected labeled source data, together with unlabeled target data, are used to learn self-training domain adaptation classifiers for the target crisis. Experimental results have shown that overall the hybrid domain adaptation classifiers perform better than the supervised classifiers learned from the original source data.

KEYWORDS
Crisis Response, Domain Adaptation, K-Nearest Neighbors, Matrix Factorization, Self-Training, Tweet Classification

INTRODUCTION

Social media is becoming a more prevalent part of our everyday life, due to the advancements in technology and virtualization. The availability of the Internet, cameras and real-time message boards at our fingertips has brought about live and parallel reporting, and witness testimonies during many events. These reports can be useful to responders and can help create awareness among the populace, especially in emergency situations (Meier, 2015; Watson, Finn, and Wadhwa, 2017). Despite the potential benefits, major response groups and organizations under-utilize these sources of information, as therein lie many administrative and technical challenges (Meier, 2013). Among the challenges, there are reliability issues associated with public and unstructured data, as well as

DOI: 10.4018/IJISCRAM.2019070101

Copyright © 2019, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
information overload issues, as millions of messages are posted during a crisis situation (Bullock, Haddow, and Coppola, 2012).

There are many recent studies that propose the use of machine learning techniques to provide automated methods for analyzing social media data to reduce the information overload (Imran et al., 2015; Beigi et al., 2016). Machine learning techniques can help transform raw data into usable information by labeling, prioritizing and structuring data, and making them beneficial to responders and to the populace in times of need (Qadir et al., 2016). However, supervised learning algorithms rely on labeled training data to build predictive models. Accurate labeling of data for an emerging crisis is both time consuming and expensive, and, hence, it is not appropriate to assume that labeled data for a current crisis will be promptly available to be used for analysis. The lack of labeled data for emerging crisis events prohibits the use of supervised learning techniques.

To address this challenge, several works proposed to use labeled data from prior “source” crises to learn supervised classifiers for a “target” crisis (Verma et al., 2011; Imran et al., 2013; Imran, Mitra, and Srivastava, 2016). However, due to the divergence of each crisis in terms of location, nature, season, etc. (Palen and Anderson 2016), the source crisis might not accurately represent the characteristics of the target crisis (Qadir et al., 2016; Imran et al., 2015). Domain adaptation techniques (Pan and Yang, 2010; Jiang, 2008) are designed to circumvent the lack of labeled target data by making use of unlabeled target data as guideposts for the readily available labeled source data. Studies in the emergency space have shown that using domain adaptation techniques, which use target unlabeled data and source labeled data together, significantly improve classification results as compared to supervised techniques that solely use labeled source data (Li et al., 2015, 2017). Unlabeled data from the target crisis becomes more abundant as the event unfolds, and it can enable the use of domain adaptation techniques during emerging or occurring crisis events.

There are several ways in which the unlabeled target data can be used with domain adaptation techniques, including parameter-based adaptation, instance-based adaptation and feature-based adaptation (Pan and Yang, 2010). In the parameter-based adaptation, the labeled source data is used together with the unlabeled target data to identify shared parameters that result in good predictions for the target data. In the instance-based adaptation, the unlabeled target data is used to identify and/or reweigh the most relevant source labeled instances with respect to the target classification task, while in feature-based adaptation, the target unlabeled data and source labeled data are used together to find a feature representation that minimizes the difference between the two domains. Relevant prior work on crisis tweet classification using domain adaptation has relied on parameter-based adaptation. Specifically, Li et al. (2017) proposed to learn weighted source and target Naïve Bayes classifiers with the iterative methods of Expectation-Maximization (EM) (Dempster, Laird, and Rubin, 1977) and Self-Training (ST) (Yarowsky, 1995) and showed that the resulting classifiers can accurately predict the target.

Mazloom et al. (2018) proposed to use a combination of two domain adaptation approaches, specifically a hybrid between feature-based adaptation and instance-based adaptation, to reduce the variation between the two domains. First, the Alternating Nonnegative Least Squares Matrix Factorization (LSNMF) in (Lin, 2007) is used on the combined source and target data, represented using binary vectors, to create a dense and reduced conceptual representation of source and target instances. Subsequently, the k-Nearest Neighbors algorithm (kNN) is used to select a subset of the source instances which are most similar to the target instances, according to the cosine similarity calculated based on the reduced common representation. Finally, the selected subset is subsequently used to learn Naïve Bayes classifiers for the target crisis.

In this study, we propose to use a combination of all three domain adaptation approaches, specifically a hybrid between feature-based adaptation, instance-based adaptation, and parameter-based adaptation to reduce the variation between the source and target domains. As in Mazloom et al. (2018), we first use the Alternating Nonnegative Least Squares Matrix Factorization (LSNMF) Lin (2007) on the combined source and target data to create a dense and reduced conceptual representation
17 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/article/a-hybrid-domain-adaptation-approach-for-identifying-crisis-relevant-tweets/235427?camid=4v1

www.igi-global.com/e-resources/library-recommendation/?id=2

Related Content

A Distributed Scenario-Based Decision Support System for Robust Decision-Making in Complex Situations
www.igi-global.com/article/distributed-scenario-based-decision-support/60613?camid=4v1a
M-Government for Emergency Notifications
[www.igi-global.com/chapter/m-government-for-emergency-notifications/207601?camid=4v1a](www.igi-global.com/chapter/m-government-for-emergency-notifications/207601?camid=4v1a)

Communicating Location and Geography in Emergency Response
[www.igi-global.com/article/communicating-location-and-geography-in-emergency-response/175673?camid=4v1a](www.igi-global.com/article/communicating-location-and-geography-in-emergency-response/175673?camid=4v1a)

Security Issues on Outlier Detection and Countermeasure for Distributed Hierarchical Wireless Sensor Networks