Classifying Text Messages for the Haiti Earthquake

Cornelia Caragea, Nathan McNeese, Anuj Jaiswal, Greg Traylor, Hyun-Woo Kim, Prasenjit Mitra, Dinghao Wu, Andrea H. Tapia, Lee Giles, Bernard J. Jansen, John Yen
College of Information Sciences and Technology
The Pennsylvania State University, University Park, PA-16801
{ccaragea, ajaiswal, hxk263, pmitra, dwu, atapia, giles, jjansen, jyen}@ist.psu.edu
{nathan.mcneese, gtraylo}@gmail.com

ABSTRACT

In case of emergencies (e.g., earthquakes, flooding), rapid responses are needed in order to address victims' requests for help. Social media used around crises involves self-organizing behavior that can produce accurate results, often in advance of official communications. This allows affected population to send tweets or text messages, and hence, make them heard. The ability to classify tweets and text messages *automatically*, together with the ability to deliver the relevant information to the appropriate personnel are essential for enabling the personnel to *timely* and *efficiently* work to address the most urgent needs, and to understand the emergency situation better. In this study, we developed a reusable information technology infrastructure, called Enhanced Messaging for the Emergency Response Sector (EMERSE). The components of EMERSE are: (i) an iPhone application; (ii) a Twitter crawler component; (iii) machine translation; and (iv) automatic message classification. While each component is important in itself and deserves a detailed analysis, in this paper we focused on the *automatic classification component*, which classifies and aggregates tweets and text messages about the Haiti disaster relief so that they can be easily accessed by non-governmental organizations, relief workers, people in Haiti, and their friends and families.

Keywords

Text message classification, abstract features, machine learning.

INTRODUCTION

As social media and more specifically, microblogging, become more integrated into the daily lives and everyday communication patterns, scholars of disasters and emergency response see challenges and hope in these practices. The existence of over four billion cell phones throughout the world, coupled with information sharing sites such as Facebook, Flikr, Twitter, and virtual universal connectivity, provide the technological basis for worldwide information collection, sharing and dissemination. These microblogging practices are often described as rich sources of timely data that may offer affected individuals and responders valuable information. According to (Vieweg, Hughes, Starbird, and Palen, 2010), microblogging is seen to have intrinsic value across responder organizations and victims because of its growing ubiquity, communications rapidity, and cross-platform accessibility. In disasters, average citizens on the ground can offer information describing the magnitude of the crisis, keeping outsiders informed as to the status on the ground.

There have been several studies on the use of social media and microblogging during crisis, including studies that have focused on the 2007 Virginia Tech shootings, the 2008 Northern Illinois University shootings, as well as the 2007 southern California wildfires (Palen, Vieweg, Liu, and Hughes, 2009; Palen and Vieweg, 2008; Vieweg, Palen, Liu, Hughes, and Sutton, 2008; Sutton, Palen, and Shklovski, 2008). For the greater part, these scholarly and journalistic accounts argue that microblogging has the potential to provide for citizen empowerment with a many-to-many information flow. Some researchers (Palen, Anderson, Mark, Martin, Sicker, Palmer, and Grunwald, 2010) argue that social media used around crises involves self-organizing behavior that can produce accurate results, often in advance of official communications.

Reviewing Statement: <to be completed by the editors> This paper has been fully double blind peer reviewed./This paper represents work in progress, an issue for discussion, a case study, best practice or other matters of interest and has been reviewed for clarity, relevance and significance.

One example of microblogging being used during crisis was the 7.0 Earthquake in Haiti. This disaster led to the mobilization of much of the world to support the relief effort, especially through novel uses of the cyberspace. Relief workers, reporters, and non-governmental organizations (NGOs) have used tweets and text messages extensively to spread and share information about the needs, events, and causalities in the Twitterworld. Regular citizens have also employed Twitter to rally others to support relief efforts. Hope140.org provides RSS feeds about tweets related to Haiti and suggests Twitter accounts to follow. Both Haitians and relief workers have used mobile phones to send text messages regarding damages, resource needs, and security-related events.

Even though the earthquake damaged much of the communication infrastructure in Haiti, Haiti's Internet connectivity was robust because most Haitian ISPs use satellite, rather than damaged undersea fiber optic cable links, to connect to Internet. To help NGOs connect to satellites, telecommunication companies (e.g., International Telecommunications Union and Trilogy International Partners) quickly set up cellular systems. Moreover, FrontLineSMS:Medic set up a free phone line, 4636, to allow affected population to send text messages, and hence, make themselves heard (Munro, 2010).

While there is useful information in these tweets and text messages, they are not well-organized to allow critical information (e.g., water, medical supply, food) to be delivered to those who need them in a timely and efficient fashion. Relief workers from different organizations, such as NGOs, military units, and government agencies, need IT support for analyzing tweets and text messages for ease of aggregation and targeted real-time broadcasting. Hence, the ability to classify tweets and text messages *automatically*, together with the ability to deliver the relevant information to the appropriate personnel are essential for enabling the personnel to timely and efficiently work to address the *most urgent needs*, and to understand the emergency situation better in the Emergency Response Sector.

Although tweets and text message classification can be performed with little or no effort by people, it still remains difficult for computers. Machine learning offers a promising approach to the design of algorithms for training computer programs to efficiently and accurately classify short text message data. Some of the main challenges in classifying such data are as follows: (i) tweets and text messages contain only a few words and, sometimes, require background information for accurate classification. For example, the message *"I live in Leogane, Route de Mellier Bongnotte #72, I need formula for my baby."* requires knowledge that *formula* refers to *baby food*. The choice of features used to encode such data is crucial for the performance of learning algorithms; (ii) tweets and text messages may belong to multiple categories (i.e., the *multi-label problem*); (iii) there may be possible errors in the manually generated labels (i.e., categories) of messages, which can impact the performance of learning algorithms; and (iv) the training (*labeled*) set is often limited in size.

Against this background, we propose and develop a reusable information technology infrastructure called EMERSE (Enhanced Messaging for the Emergency Response SEctor) that classifies and aggregates tweets and text messages about the Haiti disaster relief (originating in Haiti and elsewhere) so that they can be easily accessed by NGOs, relief workers, people in Haiti, and their friends and families.

We focused on the choice of features that are used to represent short text messages in a *multi-label setting*. We compared four types of feature representations provided as input to machine learning classifiers to accurately classify text messages from the Haiti earthquake, submitted to Ushahidi-Haiti (http://haiti.ushahidi.com) through phone, e-mail, Twitter, or web. These feature representations are obtained using: (i) a BoWs, (.e., all words in the vocabulary) (McCallum and Nigam, 1998); (ii) feature abstraction methods, that find a partition of the set of words in the vocabulary by clustering words based on the similarity between the class distributions that they induce (Silvescu, Caragea, and Honavar, 2009); (iii) feature selection methods, that select a subset of features based on some chosen criteria (Kira and Rendell, 1992); and (iv) Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan, 2003), that finds hidden topics in the data. The topic words (i.e., the words in each topic) can be seen as discriminative features.

The results of our experiments show that feature abstraction generates features that can yield better performing classifiers than those obtained by using a BoWs, features chosen by feature selection, and features as topic words output by LDA. We also discuss the insights gained from these results and suggest directions of future research to enhance the accuracy and coverage for classifying tweets and text messages for improved efficiency and coordination during the response, transition, and recovery of extreme events.

The rest of the paper is organized as follows: Section 2 discusses the related work, Section 3 describes the architecture of the EMERSE system, through an example; Section 4 presents a brief overview of feature abstraction (Silvescu et al., 2009), Relief feature selection (Kira and Rendell, 1992), and LDA (Blei et al., 2003). Section 5 provides experimental design and results, and Section 6 concludes the paper with a summary and discussion.

RELATED WORK

The problem of learning classifiers from short text messages has started to receive significant attention in the machine learning literature. Healy, Delany, and Zamolotskikh (2005), Hidalgo, Bringas, Sanz, and Garcıa (2006), and Cormack, Hidalgo, and Sanz (2007) have previously addressed the problem of identifying *spam* short messages, by employing various machine learning algorithms (such as Naive Bayes, SVM, Logistic Regression, and Decision Trees) and various feature representations (such as BoWs, BoWs augmented by statistical features, e.g., the proportion of upper case letters or punctuation in the text, orthogonal sparse word bigrams, character bigrams and trigrams). Gupta and Ratinov (2008) have employed transfer learning techniques to classify short online dialogs, by enriching the set of features using external data sources. Munro and Manning (2010) have focused on classifying medical text messages, written in Chichewa language, that were received by a clinic in Malawi, and have shown that incorporating morphological and phonological variation could improve classification performance. Furthermore, Munro (2010) has presented a brief survey about the crowdsourced translation to English of text messages written in Haitian Creole during the January 12 earthquake in Haiti. Collaborating online, people around the world were able to translate more than 40,000 messages in a short time, which led to saving hundreds of lives, and direct the food and medical aid to tens of thousands (Munro, 2010). Starbird and Stamberger (2010) introduced a Twitter hashtag syntax for reporting events related to crisis.

Unlike these works, we focused on correctly classifying text messages for the emergency response sector by determining a subset of features that are most informative for the target variable, either by selecting a subset of features from the entire vocabulary using feature selection or LDA, or by constructing abstract features using feature abstraction. In addition, our text message classification task is harder due to its *multi-label* nature (i.e., text messages may belong to multiple categories).

The topics related to emergency response (ER) form an ontology that can be applied to emergency response for a wide range of relief operations. Ontology development tools such as Protege have been widely used for developing ontology for different domains. Li, Liu, Ling, Zhan, An, Li, and Sha (2008) proposed an ontology for emergency response (ER). The top level concepts of the proposed ontology include: aftermath-handling, emergency-rescue, emergency-response, and response-preparation. Each concept is further refined by a set of subconcepts. Emergency-rescue, for instance, include medical-aid, evacuation, and victim-assistance. Turoff, Chumer, Van de Walle, and Yao (2006) have designed a dynamic emergency response management system DERMIS, and have identified the characteristics of a good ER system.

EMERSE - ENHANCED MESSAGING FOR THE EMERGENCY RESPONSE SECTOR

Before presenting the details of message classification, we first provide a brief description of our EMERSE system, which can best be described using two case scenarios: the first case involves survivors of Haiti communicating to NGOs through the system, and the second involves NGOs communicating to survivors of Haiti. Both scenarios involve classifying incoming messages and broadcasting outgoing messages to those who subscribe for them.

Scenario 1: A woman, Mrs. A, in a village in Haiti notices several of her neighbors in the same village have developed symptoms of a serious disease. She sends a short text message in French Creole through text messaging to a local phone number advertised by the relief organizations. The message is forwarded to EMERSE, which is first translated to English, then classified to the "Medical Needs" category. The name of the village in the message is extracted, and its location identified. A U.S. relief worker, Mr. B, working for an NGO for handling medical aid for the region in which the village is located has previously subscribed to the system for receiving tweets in the "Medical Needs" category for the region he is responsible for. Mr. B notices a new tweet sent by EMERSE that describes multiple villages with serious illness. Concerned about the potential to trigger a pandemic, Mr. B. uses his smart phone to forward the tweets to several nearby temporary hospitals and to the medical supply center of the NGO to alert them about the situation. Within an hour, a medical care team, consisting of doctors and nurses from multiple hospitals, carrying medical supplies obtained from the NGO, arrived at the village. Due to their prompt response, they were able to cure people in the infected villages.

Scenario 2: However, there were not enough vaccines for all the villages not infected by the disease yet. Mr. B separated all the infected villagers from those not infected. Mr. B also helped Mrs. A and other villagers with cell phones to create a Twitter account on Mr. B's laptop, and register their cell phones to receive text messages regarding the availability of medical supplies to the region. Three days later, the vaccine arrives at the medical supply center of the NGO. The worker sends a tweet to EMERSE, which is classified as "Medical Supplies" and broadcasted to those relief workers who subscribe for it. Mr. B receives the tweet, contacts the medical supply center to reserve the amount needed for the village. Mr. B then sends a tweet (in English) to the EMERSE account for the village to tell them to be in the village to receive vaccine within an hour, which is

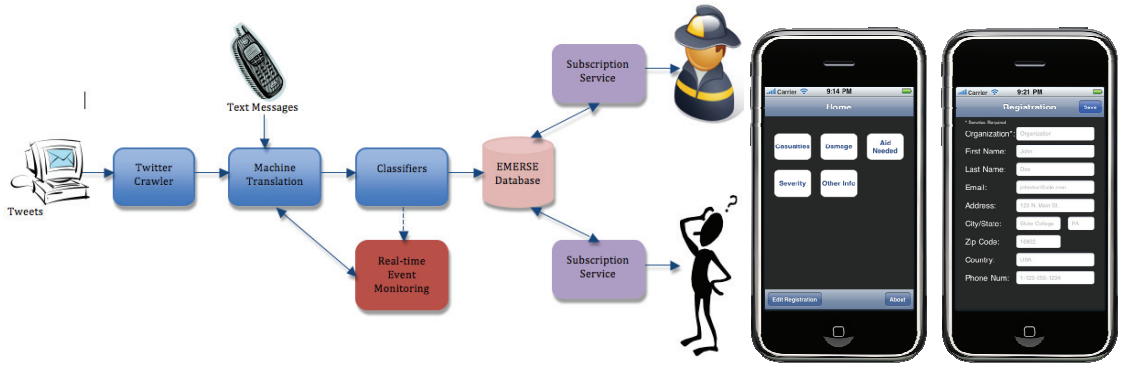

Figure 1: (a) EMERSE Architecture; (b) The iPhone Application

translated by the machine translation component to French Creole and sent to cell phones of the villages as well as nurses in nearby hospitals. One hour later, Mr. B arrives at the village with nurses and enough vaccine for the villagers. Because villagers with cell phones quickly inform other villagers, all the uninfected villagers have gathered and are ready to receive the vaccine when Mr. B arrives. A potential outbreak of a dangerous infectious disease is thus prevented because of EMERSE.

The EMERSE system architecture is presented in Figure 1(a). In addition to the *automatic classification component*, the main components of the system are as follows: (i) an iPhone application; (ii) a Twitter crawler component; and (iii) a machine translation component.

The iPhone application. We designed and implemented an iPhone application for NGOs and first responders, to assist them in recording data (situation reports, rapid assessments, etc.), that are useful in organizing emergency response. Currently, NGOs are using clipboards, pen, and paper to record observed information. Then, at the field camps, these paper reports are processed and are phoned or emailed back to the command center and distribution points (Beiser, 2010). However, newer technology is available and is being used on the ground.

When developing the application for the iPhone platform, the user interface is extremely important. The primary goal for this application was simplicity and accessibility. We would not want users to find pen and paper faster than the iPhone! Furthermore, the capacity of the iPhone to take and tag photos can be extremely useful in relaying information of a situation in a manner that words on a form cannot. Finally, geo-locating capabilities with the iPhone provide location data for each submission so that the aid can be delivered as precisely as possible. The location could be an indication of trust in the validity of the report, which is an attractive feature. The application requires registration, but once registered, it allows submission of data related to casualties, damage, aid needed, overall severity of the area, and a general notes page for anything else of consequence. Data from each page is then submitted to a server, which enables the registration and submission data to be saved remotely, and photos to be uploaded. In case of no service or internet access, the application rolls over to local storage until access is available. However, the GPS still works via satellite and is able to geo-locate data being recorded. Figure 1(b) shows the registration page and the main submission data page.

The Twitter crawler component. We designed and implemented a distributed crawl mechanism, which effectively captures tweets of most relevance to the system, e.g. tweets referring to specific crisis events. The crawler stores these tweets into a database allowing this information to be retrieved for effective analysis. The crawler utilizes the Twitter© API that retrieves relevant tweets to user-input seed keywords ("earthquake", "haiti earthquake", etc.) within a seven-day period as well as tweets that have been posted by specific users. Each API call returns at most 1000 tweets and auxiliary metadata (e.g., creation time, geo-location, tweet id, user id, etc.) in JSON format, which is written to a file on disk. This file is parsed and each tweet is written to a postGRES database such that each tweet having the same id (returned by Twitter) appears only once. Since Twitter assigns each new retweet with a different id, these are stored as unique database entries provided retweets are returned when querying for a specific keyword. Furthermore, once the crawler downloads tweets for specific keywords for the first time, it updates a timestamp column so that each request for more tweets is limited only to the current date. (For the first request, the returned tweets may be for at most a seven-day duration). Due to Twitter call limitations, the crawler typically sends about 150 keyword requests per hour.

The machine translation component. We utilize Google AJAX Language API[1] and Google Haitian Creole translator[2] to translate tweets in French Creole and other languages. We observed posts in French, German,

[1] http://code.google.com/apis/ajaxlanguage/documentation/

Japanese, Dutch, etc. related to the Haitian catastrophe. We tested the API with a sample of existing tweets in French Creole from twitter and verified that the quality of the translations is high especially for the purposes of our learning algorithms, in which the linguistic and grammatical errors like the order of verbs, etc. do not matter. The Google API translates documents up to 1MB in size and can be used to translate 1GB translations per year. This equates to more than 7 million tweets. Google's quota has been implemented to prevent unethical uses and denial of service attacks. Thus, we remain confident that in the unforeseen case that the volume of tweets increases drastically, we will contact Google to allow us to allocate a higher quota and should be able to procure these additional services.

Next, we present details of the *automatic classification component* of EMERSE.

LEARNING AND CLASSIFICATION

The supervised learning problem (Duda, Hart, and Stork, 2001) can be formally defined as follows. Given: (i) an *independent and identically distributed* (*iid*) data set $\mathcal{D}$ of labeled examples $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$, $\mathbf{x}_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, where $\mathcal{X}$ denotes a vocabulary of words and $\mathcal{Y}$ the set of all possible class labels; (ii) a hypothesis class $\mathcal{H}$ representing the set of all possible hypotheses that can be learned; and (iii) a performance criterion P (e.g., accuracy), then a learning algorithm L outputs a hypothesis $h \in \mathcal{H}$ (i.e., a classifier) that optimizes P. During classification, the task of the classifier h is to accurately assign a new example $\mathbf{x}_{test}$ to a class label $y \in \mathcal{Y}$ (see Figure 2).

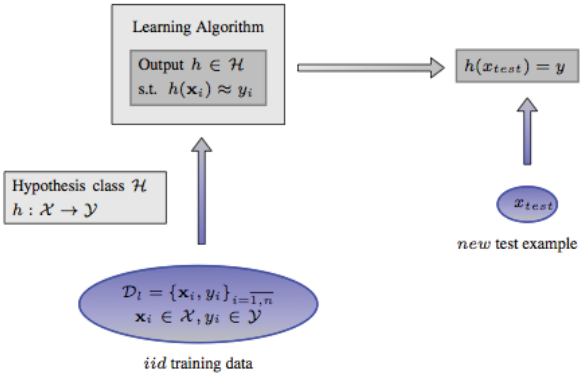

Figure 2: Supervised Learning and Classification.

In the remaining of this section, we describe four methods, which produce feature representations that are used as input to machine learning algorithms: (1) the BoWs approach; (2) feature abstraction; (3) feature selection; and (4) Latent Dirichlet Allocation (LDA). The motivation for choosing these four methods is that they were successfully used for text document modeling and classification. The "bag of words" approach is commonly used for text classification; feature abstraction methods reduce a model input size by grouping "similar" features into clusters of features; feature selection methods attempt to remove irrelevant or redundant features in order to improve classification performance; and LDA models are designed to discover clusters of semantically related words that co-occur in a collection of documents.

The BoWs approach. The BoWs approach is widely used by the machine learning community for many classification tasks (see, for instance, (McCallum and Nigam, 1998)). Each example (i.e., a text message) is drawn from a multinomial distribution of words from a vocabulary, and the number of independent trials is equal to the length of the example. Hence, each example is represented by its words (used as features in the classification model), which are assumed to be independent given the class variable.

Feature Abstraction. Feature abstraction methods are potentially successful techniques for producing appropriate features for classification (Silvescu et al., 2009). They reduce the classifier input size by grouping "similar" features to generate *abstract features* (also called abstractions). Silvescu et al. (2009) proposed an approach to simplifying the data representation used by a learner by grouping features based on the Jensen-

[2] http://code.google.com/p/translator-ht/

Shannon divergence (Cover and Thomas, 1991) that result in minimal reduction in the mutual information between features and the class variable.

Specifically, they used hierarchical agglomerative clustering to group the most "similar" features at each step of the algorithm, based on the similarity between the conditional distributions of the class variable given the features. The most "similar" features are identified as those that have the smallest Jensen-Shannon divergence between the conditional distributions of the class that the features induce. As an effect, abstract features that are predictive of the class variable are obtained. An example of an abstract feature can be "food", which is more general than the specific features "rice" and "formula" (i.e., baby food). The abstract feature is identified by the group {*rice, formula*}.

Silvescu et al. (2009) have shown that *abstraction* reduces the model input size and helps improve the statistical estimates of complex models (especially when data are sparse) by reducing the number of parameters to be estimated from data. In this study, we have applied the feature abstraction approach of Silvescu et al. to generate the abstract feature representation.

Feature Selection. Feature selection methods attempt to remove redundant or irrelevant features in order to improve classification performance of learning algorithms (Guyon and Elisseeff, 2003). Feature selection selects a subset of the available features based on some chosen criteria, and can substantially reduce the number of model parameters. Kira and Rendell (1992) proposed an algorithm for feature selection, called Relief, which is not heuristic-based, is robust to noise and to interaction among features.

Relief is a weight-based algorithm. At each step, Relief samples from the training data an instance x, and determines x's *near-hit* (the closest instance from the same class as x) and *near-miss* (the closest instance from the opposite class of x) in the training data, by using p-dimensional Euclidian distance. A feature weight vector is updated for each such triplet to determine the relevance of all features to the class variable. The algorithm terminates after k steps and returns those features whose relevance level is above some user-specified threshold. We have used the Relief algorithm to select a subset of features that are predictive to the class variable.

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) is an unsupervised method for detecting hidden topics in the data proposed by Blei et al. (2003). LDA is a generative probabilistic model of a collection of documents, which has been successfully used to perform dimensionality reduction for text classification (Wei and Croft, 2009), where documents are multiple paragraphs and pages in length.

LDA models each document in a collection as a mixture of topics (drawn from a conjugate Dirichlet prior), and each topic as a distribution over words in the vocabulary. The topic distribution of a document can be seen as a lower dimensional representation of the document (where the dimensionality is equal to the number of topics). Furthermore, the union of the words with high probability in each topic can be seen as a set of discriminative features for the collection of documents. We have used these words to generate the topic words feature representation.

EXPERIMENTS AND RESULTS

Ushahidi Text Message Data Set. The data set used in our experiments is the Ushahidi data set (http://haiti.ushahidi.com/), which consists of 3,598 text messages from the Haiti earthquake. We used a subset of 2,116 text messages of the Ushahidi data set, for which the English translation is available. While text messages are available in both Haitian Creole and English languages, we used only the English version, as Munro and Manning (2010) found no significant improvement from one language to another on a similar task. The messages have been manually labeled into 10 categories: (1) *medical emergency*; (2) *people trapped*; (3) *food shortage*; (4) *water shortage*; (5) *water sanitation*; (6) *shelter needed*; (7) *collapsed structure*; (8) *food distribution*; (9) *hospital/clinic services*; and (10) *person news*. A message may belong to multiple categories. For example, the message *"Good evening ONG, I'm very happy for the aid you're giving to the people, I thank you. But in my zone that's to say Lamenten 54 Rue St Juste we need shelter and food."* belongs to both *shelter needed* and *food shortage* categories. In experiments, we compared the predicted labels produced by machine learning algorithms against the manual labels generated by people. Note that we could have employed the machine translation component of our EMERSE system to translate messages from Haitian Creole to English. However, as already stated, our main focus here is on the automatic classification component of EMERSE.

Experimental Design. Next, we describe two different ways for performing classification: classification by keywords and classification by Support Vector Machine (SVM), which is a machine learning algorithm.

Classification by keywords is a simple technique for automatic classification. For each class (or category), we identify a set of keywords as follows: we sort the words from the vocabulary based on their frequency of occurrences in the messages that belong to that particular class. The top k most frequent words from each class are considered as the keywords for each class. We ensure that there are no overlapping keywords among classes. During classification, each example is assigned to a particular class if it contains at least one keyword from the set of keywords for that class.

Classification by SVMs. SVM classifiers are among the most effective machine learning algorithms for many complex binary classification problems (Burges, 1998). Given a set of labeled inputs $(\mathbf{x}_i,y_i)_{i=1,\ldots,l}$, $\mathbf{x}_i \in \mathbf{R}^d$ and $y_i \in \{-1,+1\}$, learning an SVM classifier is equivalent to learning a decision function $f(\mathbf{x})$ whose sign represents the class assigned to input $\mathbf{x}$. This can be achieved by solving a dual quadratic optimization problem. During classification, an unlabeled input $\mathbf{x}_{test}$ is classified based on the sign of the decision function, $sign(f(\mathbf{x}_{test}))$ (i.e., if $f(\mathbf{x}_{test})>0$, then $\mathbf{x}_{test}$ is assigned to the positive class; otherwise, $\mathbf{x}_{test}$ is assigned to the negative class) (Vapnik, 1998).

Our experiments using SVMs are designed to explore what feature representations of short text messages, provided as input to SVMs, result in best classification performance. We used four types of feature representations to enable learning SVM classifiers on the Ushahidi text message data set:

- a bag of words representation, i.e., all words in the vocabulary. After stemming and removing stop words, and words with document frequency less than 3 (i.e., words that occur in less than 3 messages), the vocabulary size is 1525 (McCallum and Nigam, 1998);
- a bag of m words chosen using the Relief feature selection method (FS) (Kira and Rendell, 1992);
- a bag of m abstractions over all words in the vocabulary, i.e., an m-size partition of the vocabulary obtained by grouping words into m abstract terms based on the similarity between the class distributions that they induce (FA) (Silvescu et al., 2009);
- a bag of m topic words output by Latent Dirichlet Allocation (LDA) as the top 20 words from k topics (the number of topic words m is bounded by $20\times k$) (TW) (Blei et al., 2003).

In our experiments, we used the WEKA implementation (Hall, Frank, Holmes, Pfahringer, Reutemann, and Witten, 2009) of SVM with the default parameters, and the MALLET implementation (McCallum, 2002) of LDA. The LDA parameters are set to default, except for the number of iterations of Gibbs sampling, which is set to 3,000, and the random seed, which is set to 1. The number of topics k is set to 9 (chosen to be close to the number of categories in the data set). This results in m=165 topic words. Hence, we trained classifiers for m=165 for all of the above feature representations. In the case of feature abstraction, the 165-size partition of the vocabulary produces classifiers that use smaller number of "features" compared to the BoWs representation, (i.e., 1525 words) and at the same time, the model compression is not very stringent so as to lose important information in the data through *abstraction*.

Because a text message may belong to one or more categories, in the classification component of the EMERSE system, once a text message is received, it is desirable to assign it to all the categories it belongs to, and hence, deliver it to the appropriate departments, in a timely and efficient fashion. The classification problem in which examples belong to multiple categories is referred to as the *multi-label classification* problem. To solve this, we considered a natural decomposition into 10 "one vs. others" binary classification problems, and trained 10 binary classifiers, one for each category. The categories assigned to a text message are those predicted as positive by each of the 10 classifiers.

For all experiments, we report the average F1 measure obtained in a 5-fold cross-validation experiment. *F1* measure is the harmonic mean of precision (P) and recall (R), i.e., $F1=\frac{2PR}{P+R}$, $P=\frac{TP}{TP+FP}$ and $R=\frac{TP}{TP+FN}$. TP, FP, and FN denote true positives, false positives, and false negatives, respectively. We present results for each individual binary classifier and for the *multi-label classifier*. For the latter, precision and recall are computed as follows: $P=\frac{\sum_{i=1}^{10} TP_i}{\sum_{i=1}^{10}(TP_i+FP_i)}$ and $R=\frac{\sum_{i=1}^{10} TP_i}{\sum_{i=1}^{10}(TP_i+FN_i)}$, where TP_i, FP_i, and FN_i denote true positives, false positives, and false negatives, respectively, for the i^{th} classifier.

Results. Table 1 shows the comparison of average F1 measure (along with 95% confidence intervals) obtained using the classification by keywords and the classification by binary SVMs trained on the Ushahidi text message data set for each of the ten categories. The feature representations used to train the SVM classifiers are as follows: (i) BoWs; (ii) abstractions used as "features" in the classification model, which are obtained by feature

abstraction (FA); (iii) features selected by Relief feature selection (FS); and (iv) topic words, output by LDA (TW). Table 2 shows similar results for the *multi-label classification* setting.

Class	Classification by Keywords	Support Vector Machines			
		BoW	FA	FS	TW
medical emergency	0.20	**0.29 ± 0.06**	0.27 ± 0.08	0.12 ± 0.07	0.11 ± 0.05
people trapped	0.52	0.68 ± 0.11	**0.74 ± 0.09**	0.64 ± 0.14	0.62 ± 0.23
food shortage	0.71	0.71 ± 0.02	**0.73 ± 0.03**	0.71 ± 0.06	0.72 ± 0.07
water shortage	0.63	0.66 ± 0.03	**0.67 ± 0.02**	0.63 ± 0.04	0.65 ± 0.03
water sanitation	0.83	0.91 ± 0.01	0.94 ± 0.01	**0.96 ± 0.01**	0.95 ± 0.01
shelter needed	**0.57**	0.52 ± 0.02	0.52 ± 0.05	0.44 ± 0.07	0.48 ± 0.04
collapsed structure	0.38	**0.42 ± 0.08**	0.33 ± 0.15	0.31 ± 0.16	0.39 ± 0.20
food distribution	0.24	**0.27 ± 0.05**	**0.27 ± 0.03**	0.18 ± 0.07	0.17 ± 0.09
hospital/clinic services	0.55	0.56 ± 0.04	**0.59 ± 0.06**	0.47 ± 0.08	0.51 ± 0.05
person news	0.28	0.55 ± 0.06	**0.59 ± 0.04**	0.39 ± 0.10	0.45 ± 0.04

Table 1: Comparison of average F1 Measure (with 95% confidence intervals) obtained in 5-fold cross-validation experiments using classification by keywords and classification by binary SVM classifiers trained on the Ushahidi text message data set for each of the ten classes. The feature representations used to train the classifiers are as follows: (i) "bag of words" (BoW); (ii) abstractions used as "features" in the classification model, which are obtained by feature abstraction (FA); (iii) features selected by Relief feature selection (FS); and (iv) topic words, output by LDA (TW).

	Classification by Keywords	Support Vector Machines			
		BoW	FA	FS	TW
Multi-label classification	0.47	0.56 ± 0.01	**0.59 ± 0.01**	0.54 ± 0.03	0.56 ± 0.03

Table 2: Comparison of average F1 Measure (with 95% confidence intervals) obtained in 5-fold cross-validation experiments using classification by keywords and classification by binary SVM classifiers trained on the Ushahidi text message data set for the *multi-label* classification. The feature representations used to train the classifiers are as in Table 1.

As can be seen from Table 1 and 2, classification by keywords, despite its simplicity, often produces results that are comparable with those obtained by using SVM classifiers. Furthermore, feature abstraction (FA) significantly outperforms BoWs for most of the categories from the Ushahidi text message data set, using SVM classifiers. This suggests that FA can help minimize *overfitting* (through parameter smoothing). For few categories, for example *shelter needed*, FA-based SVM matches the performance of BoW-based SVM with substantially smaller number of features, i.e., 165 and 1525 features are used for training FA-based SVM and BoW-based SVM, respectively.

Compared to feature selection (FS) and topic words (TW), FA significantly outperforms both of them for the same number of features used in the classification model, on all categories except *water sanitation*. Although topic models have been successfully applied to documents that are multiple paragraphs and pages in length, we found that they do not work very well when applied to short text messages.

It is interesting to note that the performance of binary SVM classifiers varies significantly from one category to another, with some categories being harder to classify than others, e.g., *medical emergency* vs. *water sanitation*. This could be due to not very informative features that are used to represent the data. Hence, design of better suited feature representations for such categories will be studied in future work.

SUMMARY AND DISCUSSION

Summary. In this study, we presented a system, called Enhanced Messaging for the Emergency Response SEctor (EMERSE), whose main components are: (i) automatic message classification; (ii) an iPhone application; (iii) a Twitter crawler component; and (iv) machine translation. While each component is important in itself and deserves a detailed analysis, in this paper we focused on the automatic classification component with the goal to correctly classify messages and deliver them to the appropriate departments in a timely and efficient fashion. We compared four types of feature representations for learning SVM classifiers to *accurately* classify text messages about the Haiti disaster relief (originating in Haiti and elsewhere) so that they can be easily accessed by NGOs, other relief workers, people in Haiti, and their friends and families. These feature representations are: BoWs, abstract features (or abstractions), features selected using feature selection, and topic words output by LDA.

The results of our experiments on the Ushahidi text message data set show that using *abstract features* makes it possible to construct predictive models that use significantly smaller number of features than those obtained using a BoWs representation. The resulting models are competitive with, and often significantly outperform those that use the BoWs feature representation. Moreover, *abstract features* yield better performing models than features selected by Relief feature selection, and than topic words extracted using LDA.

Discussion. Even though Ushahidi has deployed an online report system (ushahidi.com) for Haiti, the proposed EMERSE system offers four important benefits not provided by Ushahidi. First, EMERSE will automatically classify tweets and text messages into topic, whereas Ushahidi collects reports with broad category information provided by the reporter. Second, EMERSE will also automatically geo-locate tweets and text messages, whereas Ushahidi relies on the reporter to provide the geo-location information. Third, in EMERSE, tweets and text messages are aggregated by topic and region to better understand how the needs of Haiti differ by regions and how they change over time. The automatic aggregation also helps to verify reports. A large number of similar reports by different people are more likely to be true. Finally, EMERSE will provide tweet broadcast and GeoRSS subscription by topics or region, whereas Ushahidi only allows reports to be downloaded.

In learning from real-world text message data, other challenges may be encountered, hence making the learning problem harder. We point out one of the most important challenges, i.e., possible errors in manually labeling text messages, and provide potential solutions that will be addressed in future work. As an example, the text message *"We in Canada turjo quote, we need food, water and tents. count on your participation"* belongs to *Food distribution*. However, this example is very similar to *"Good evening ONG, I'm very happy for the aid you're giving to the people, I thank you. But in my zone that's to say Lamenten 54 Rue St Juste we need shelter and food."*, which belongs to *Food Shortage*. In future work, we plan to create a new category that will contain examples from both *Food distribution* and *Food Shortage*.

Furthermore, possible errors in labeling may occur due to the presence of general terms in a text message. For example, the text message *"We need help at Mahotiere 79. Since the catastrophe, we have not seen anyone from the government"* is labeled as *Food distribution* and *Water sanitation* in the Ushahidi data set. However, there is no indication of the type of help needed. For example, people at Mahotiere 79 may need medical assistance or shelter. To distribute this message to food and water departments may be very inefficient if the people have other more urgent needs. Instead, we propose to use a general category, which consists of these types of messages. Hence, the general department can *efficiently* determine what the needs are and act accordingly.

Other future directions. Further research may also include: (i) Exploration of other types of abstraction based on semantically related words; (ii) Design of an emergency response ontology (expressed in OWL), which will be made easily accessible through Protege ontology library; (iii) Classification of tweets about Haiti, provided by Twitter. Here, because large amounts of unlabeled tweets are available, we will employ semi-supervised learning techniques that can incorporate information available in the unlabeled data to learn more robust classifiers; (iv) Details and analysis of the other components of EMERSE, e.g., the iPhone application, which assists NGOs and first responders in recording data that are useful in organizing emergency response.

Acknowledgements. This research is partially supported by NSF RAPID grant IIS 1026763 and by the U.S. Department of Homeland Security under Award #: 2009-ST-061-CI0001. 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 or the U.S. Department of Homeland Security. We would like to acknowledge Ushahidi for making the data available for this research, and NetHOPE for general interests about this research. We also wish to thank our anonymous reviewers for their constructive comments, which helped improve the presentation of this paper.

REFERENCES

1. Beiser, V. (2010, 19 Apr. 2010). Organizing Armageddon: What we learned from the Haiti earthquake. *Wired News*.
2. Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993-1022.
3. Burges, C. (1998). A tutorial on Support Vector Machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 121-167.
4. Cormack, G., Gomez Hidalgo, J., & Sánz, E. (2007). *Spam filtering for short messages.* Paper presented at the Proceedings of the 16th ACM CIKM
5. Cover, T., & Thomas, J. (1991). *Elements of Information Theory*: John Wiley.

6. Duda, R., Hart, P., & Stork, D. (2001). *Pattern Classification (2nd Edition)*: Wiley-Interscience.
7. Gomez Hidalgo, J., Bringas, G., Sánz, E., & Garćıa, F. (2006). *Content-based SMS spam filtering.* Paper presented at the DocEng '06: Proceedings of the 2006 ACM symposium on Document engineering.
8. Gupta, R., & Ratinov, L. Text categorization with knowledge transfer from heterogeneous data sources. *2008*. Paper presented at the Proceedings of AAAI 2008.
9. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157-1182.
10. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations, 11*.
11. Healy, M., Delany, S., & Zamolotskikh, A. (2005). An assessment of case base reasoning for short text message classification.
12. Kira, K., & Rendell, A. (1992). *The feature selection problem: Traditional methods and a new algorithm.* Paper presented at the Proceedings of AAAI, San Jose, CA.
13. Li, X., Liu, G., Ling, A., Zhan, J., An, N., Li, L., et al. (2008). *Building a practical ontology for emergency response systems.* Paper presented at the Proceedings of IEEE International Conference on CSSE '08.
14. McCallum, A. (2002). Mallet: A machine learning for language toolkit.
15. McCallum, A., & Nigam, K. (1998). A comparison of event models for Naive Bayes text classification.
16. Munro, R. (2010). *Crowd-sourced translation for emergency response in Haiti: the global collaboration of local knowledge*. Paper presented at the Relief 2.0 in Haiti.
17. Munro, R., & Manning, C. (2010). *Subword variation in text message classification*. Paper presented at the Human Language Technologies: The Annual Conference of the North American Chapter of the ACL.
18. Palen, L., Anderson, M., Mark, G., Martin, J., Sicker, D., Palmer, M., et al. (2010). *A vision for technology-mediated support for public participation and assistance in mass emergencies and disasters*. Paper presented at the ACM and British Computing Societys 2010 Conference on Visions of Computer Science.
19. Palen, L., & Vieweg, S. (2008). *The emergence of online wide scale interaction in unexpected events.* Paper presented at the Proceedings of CSCW 2008.
20. Palen, L., Vieweg, S., Liu, S., & Hughes, A. (2009). Crisis in a networked world: Features of computer-mediated communication in the April 16, 2007 Virginia Tech event. *Social Science Computer Review Special Issue on E-Social Science*.
21. Silvescu, A., Caragea, C., & Honavar, V. (2009). *Combining super-structuring and abstraction on sequence classification*. Paper presented at the ICDM.
22. Starbird, K., & Stamberger, J. (2010). *Tweak the tweet: Leveraging microblogging proliferation with a prescrip- tive syntax to support citizen reporting.* Paper presented at the Proceedings of the 7th Intl. ISCRAM Conf. '10.
23. Sutton, J., Palen, L., & Shklovski, I. (2008). *Backchannels on the front lines: Emergent use of social media in the 2007 Southern California fires.* Paper presented at the Proceedings of ISCRAM '08.
24. Turoff, M., Chumer, M., Van de Walle, B., & Yao, X. (2004). Design of a dynamic emergency response management information system (DERMIS). *Journal of Information Technology Theory and Application*.
25. Vapnik, V. (1998). *Statistical Learning Theory*. NY: John Wiley & Sons.
26. Vieweg, S., Hughes, A., Starbird, K., & Palen, L. (2010). *Microblogging during two natural hazards events: What twitter may contribute to situational awareness.* Paper presented at the Proceedings of the ACM 2010 Conference on Computer Human Interaction.
27. Vieweg, S., Palen, L., Liu, S., Hughes, A., & Sutton, J. (2008). *Collective intelligence in disaster: Examination of the phenomenon in the aftermath of the 2007 Virginia Tech shooting.* Paper presented at the Proceedings of ISCRAM '08.
28. Wei, X., & Croft, W. (2009). *LDA-based document models for ad-hoc retrieval.* Paper presented at the Proceedings of ACM SIGIR.