Early Detection of Heterogeneous Disaster Events Using Social Media

Viktor Pekar*
University of Birmingham Business School, University of Birmingham, Birmingham, B15 2TY, United Kingdom. E-mail: v.pekar@gmail.com

Jane Binner
University of Birmingham Business School, University of Birmingham, Birmingham, B15 2TY, United Kingdom. E-mail: j.m.binner@bham.ac.uk

Hossein Najafi
Computer Science and Information Systems, University of Wisconsin, River Falls, WI. E-mail: hossein.najafi@uwrf.edu

Chris Hale
Electronic Systems Lab, Georgia Tech Research Institute, Dayton, OH. E-mail: chris.hale@gtri.gatech.edu

Vincent Schmidt†
Air Force Research Laboratory, Dayton, OH. E-mail: vince@vincentive.org

This article addresses the problem of detecting crisis-related messages on social media, in order to improve the situational awareness of emergency services. Previous work focused on developing machine-learning classifiers restricted to specific disasters, such as storms or wildfires. We investigate for the first time methods to detect such messages where the type of the crisis is not known in advance, that is, the data are highly heterogeneous. Data heterogeneity causes significant difficulties for learning algorithms to generalize and accurately label incoming data. Our main contributions are as follows. First, we evaluate the extent of this problem in the context of disaster management, finding that the performance of traditional learners drops by up to 40% when trained and tested on heterogeneous data vis-á-vis homogeneous data. Then, in order to overcome data heterogeneity, we propose a new ensemble learning method, and found this to perform on a par with the Gradient Boosting and AdaBoost ensemble learners. The methods are studied on a benchmark data set comprising 26 disaster events and four classification problems: detection of relevant messages, informative messages, eyewitness reports, and topical classification of messages. Finally, in a case study, we evaluate the proposed methods on a real-world data set to assess its practical value.

Introduction

Early acquisition of situational awareness is an important measure for mitigating casualties and infrastructure damage caused by natural and man-made disasters. The present-day ubiquity of mobile devices has meant that during a mass crisis, social media are often the first to publish eyewitness reports on the events as they unfold. Social media are thus currently viewed as a major source of information for first responders that can make them better equipped to detect disasters at early stages, monitor their development, and coordinate planning of recovery operations.

Additional Supporting Information may be found in the online version of this article.

*Correspondence address: Business School, Aston University, Birmingham B4 7ET, United Kingdom. E-mail: v.pekar@aston.ac.uk

†USAF DISTRIBUTION STATEMENT A. Approved for public release: distribution is unlimited. 88ABW Cleared 1/30/2016; 88ABW-2016-6141.

Received September 13, 2017; revised August 10, 2018; accepted January 17, 2019

 • Published online Month 00, 2018 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.24208

Today, the value of the information posted on social media is widely recognized by humanitarian officials. Real-world examples include the American Red Cross, whose Digital Operations Center for Humanitarian Relief uses a social media monitoring system to track potential emergency reports; the Australian Red Cross, who use a computer system to filter spam and categorize social media posts into event types; ResilienceDirect, a newly established communication platform that enables cooperation between all UK emergency services via integrating evidence collected from various sources, including social media.

Driven by this goal, researchers attempted to find solutions to the problem of interpreting textual signals about disaster events within a variety of paradigms, such as knowledge management (Chua, 2007; Yates & Paquette, 2011) and content analysis (Choo & Nadarajah, 2014; Heverin & Zach, 2012). Message classification methods based on machine learning attract particular attention due to their ability to automate the process of analytical model building and adapt to the changing nature of data without human intervention, which are important properties in the context of disaster management, where very large amounts of data need to be sifted through to detect very specific types of messages. These methods have been successfully implemented in a number of real-world systems for disaster monitoring (for example, Imran, Castillo, Lucas, Meier, & Vieweg, 2014). Previous studies on machine-learning approaches have shown that if the message classification task is limited to a narrow domain such as floods, earthquakes, or tornadoes, relevant messages can be detected with a reasonably high degree of accuracy (for example, Caragea, Silvescu, & Tapia, 2016; Imran, Elbassuoni, Castillo, Diaz, & Meier, 2013; Musaev, Wang, Cho, & Pu, 2014). However, emergency events tend to differ substantially in terms of their causes, temporal and geographical spread, impacted targets, and the nature of damage; a specific event may combine characteristics of multiple disaster types. It is much more practical to have a classification method that can cover the widest possible range of disaster types in order to give first responders and emergency services personnel confidence that disasters with some previously unseen characteristics would be recognized by the alerting system.

This article addresses the task of recognizing reports on mass emergencies unrestricted to a particular type, which could include both natural disasters such as hurricanes, floods, and storms, as well as manmade ones such as explosions, collisions, and shootings. This is a nontrivial problem, as the data is nonhomogeneous: the classifier is trained and evaluated on data covering different emergency types; each characterized by its own vocabulary and correspondingly different classification feature distributions. Our main contributions are a new method for message classification based on ensemble classification specifically suited to the task of detecting disaster events that were unseen at the training stage, its comparative performance evaluation with traditional" base" classifiers, and other ensemble classifiers. The evaluation was conducted on four different classification tasks and under three application scenarios that were studied in previous research.

Literature Review

The recent growth of online communications has led to increased practical interest in automatic processing of short text messages, such as social media posts, instant messages, and online chat logs, in order to detect particular kinds of messages. A popular direction of work is concerned with detection of new events in a stream of messages; some of these approaches have been applied to detecting mass emergency events. Such methods primarily rely on detecting "bursty" keywords (Marcus et al., 2011), that is, keywords whose frequency increases sharply within a short time window, or bursty message clusters (Schmidt & Binner, 2015). However, bursty keywords, taken out of context, are often ambiguous, and may be related not only to new events, but also recurring events and even nonevents. To identify the most useful keywords among those with a high burstiness score, Becker, Naaman, and Gravano (2011) used a domain-independent text classifier.

Domain-specific methods generally have greater accuracy than domain-independent ones, and previous work specifically on emergency event detection was concerned with developing domain text classifiers based on machine learning and operating on features extracted from the entire message. Most of this work dealt with one particular type of crisis, such as earthquakes (Caragea et al., 2011), landslides (Musaev et al., 2014), floods (Caragea et al., 2016), or hurricanes (Fan, Mostafavi, Gupta, & Zhang, 2018).

A number of studies aimed to develop classifiers that would be applied to more than one type of disaster. Verma et al. (2011) conducted experiments on how well a classifier trained on one type of emergency would perform on messages representing a different emergency type. They ran all pairwise comparisons between four data sets, which represented two flood events, one earthquake, and one wildfire, and found that testing on an emergency type other than the one used for training results in much poorer classification accuracy; the F-measure ranging between 29% and 83%, depending on a specific pair. Similarly, Imran et al.'s (2013) study showed that there is a significant loss of accuracy when a model that is trained on one crisis (2011 Joplin tornado) is used to classify messages describing another crisis (2012 Hurricane Sandy), despite the apparent similarity between the two types of crises. Ashktorab, Brown, Nandi, and Culotta (2014) trained one generic classifier on data from 12 different emergency events, achieving an F-measure between 50% and 65%, depending on the learning method; however, the evaluation was done by randomly splitting all the data into test and training sets, that is, the training and test data contained data representing different disasters in similar proportions.

Several studies looked at methods to adapt a machine-learning classifier trained on one type of a disaster (the source domain) to some other one (the target domain), using a set of labeled tweets from the source domain and a set of unlabeled tweets representing the target domain. Such methods are seen as a solution for situations where labeled data for a particular domain are hard to obtain. Using data on the 2012 Hurricane

DOI: 10.1002/asi

Sandy as source and the 2013 Boston Bombings as target, Li et al. (2015) found that the auROC value increased considerably for the tasks of identifying crisis-related tweets when a small amount of labeled data for the target domain was supplemented with unlabeled target data. Addressing the same problem of the lack of labeled data, Imran, Mitra, and Srivastava (2016) conducted experiments with reusing labels from the source domain to classify target-domain tweets, but could not establish that this cross-domain transfer helps classification accuracy.

It should be pointed out that direct comparison between previous approaches is problematic, because somewhat different classification tasks were used. For example, Li et al. (2015) and Nguyen et al. (2017) classified messages into related and unrelated to a disaster, Caragea et al. (2016) and Derczynski, Meesters, Bontcheva, and Maynard (2018) into "informative" and "noninformative," Ashktorab et al. (2014) into those that report damage and those that do not, Verma et al. (2011) into those that contribute to situational awareness and those that do not, and Burel and Alani (2018) classified messages into multiple topical categories such as affected individuals, infrastructures and utilities, donations and volunteer, caution and advice.

Data Heterogeneity

Data heterogeneity affects many large-scale machine-learning applications (Duan, Clancy, & Szczesniak, 2016). It occurs in situations where both training and test data are drawn from multiple data sources, each characterized by its own feature distributions, which ultimately creates problems for the learning algorithm to generalize. The problem of detecting disaster-related messages independently of the disaster type that we aim to solve is an example of such a situation: Messages relating to different types of disasters tend to have different vocabularies and hence different distributions of classification features.

The efficacy of a single classifier on such heterogeneous data is often poor. One effective approach to learning from heterogeneous data is ensemble classification (see Dietterich, 2000). The basic idea behind ensemble learning is to attempt to divide the data into homogeneous subsets—by finding an underlying structure in the set of features or instances—and use multiple classification models ("weak learners") trained on different subsets to capture the diverse aspects of the data. The weak learner models are then combined to obtain a new, stronger classifier that outperforms the original ones when used separately. Ensemble methods have been widely used in many predictive learning problems to improve performance on heterogeneous data (for example, Ballings & den Poel, 2015). They have also been shown to compare favorably with traditional classification methods when applied to classification of short text messages (for example, Hagen, Potthast, Büchner, & Stein, 2015; Tuarob, Tucker, Salathe, & Ram, 2014).

Among the most popular algorithms for ensemble learning are Adaptive Boosting, Gradient Boosting (AdaBoost), and Random Forest, which are described next.

AdaBoost

AdaBoost (Freund & Schapire, 1996) uses the whole training data set to successively train a series of weak learners, such as decision stumps. After one classifier is trained, the algorithm identifies the most difficult instances and computes their weights to exaggerate their effect on the training of the next classifier. The objective of this step is to correctly classify the misclassified instances by the next classifier. Initially all instances have the same weight, and hence have the same impact on training of the initial model. After each iteration, the weights of misclassified instances are adjusted, while the weights of correctly classified instances are decreased. Furthermore, each classifier is assigned a weight based on its overall accuracy. During the testing phase, the output labels and the weights of the classifiers are considered to produce a weighted average vote across the weak classifiers.

Gradient Boosting Classifier (GBC)

Gradient Boosting (Friedman, 2001) is a gradient descent algorithm, which, similar to other boosting methods, operates by consecutive training of weak classifiers that collectively would form a strong classifier. This is accomplished by training successive classification models on the residuals of the previous model, computed from errors it made. With each training round, Gradient Boosting improves the previous model by adding to it a new model that is trained only on the residuals, thus gradually correcting errors made by previous models.

Random Forest

The Random Forest algorithm (Breiman, 2001) uses a large number of weak learners, usually deep decision trees, as building blocks to form a generalized classification model. When training one weak learner, the algorithm starts by drawing a random sample of training instances, with replacement (that is, allowing the instance to be present in multiple subsets). In addition, the selected instances are represented with a random subset of features, in order to decorrelate the classification models and reduce the variance of their output. At the testing stage, the majority of the classifier votes are output as the eventual class label.

Disaster-Based Ensembles

Ensemble methods attempt to overcome the heterogeneity in the data by finding subsets of instances that are characterized by similar feature distributions. In the context of identifying disaster-related messages, training data can be already provided with labels that indicate the disaster type of each message. We investigate the idea that the subsets of the data corresponding to these labels form a suitable structure that can be used by an ensemble classifier.

We create a classifier ensemble through dividing the training instances by disaster type and training one classifier specific to each type, using the same learning algorithm. Each of the classifiers is thus expected to be more effective at classifying just its own disaster type, than a classifier trained on other types or a generic classifier. Test

DOI: 10.1002/asi

instances representing an unknown disaster would then be classified more effectively by certain classifiers compared to others. This is because the unknown disaster will be more similar to some of the disaster types observed during training than others.

The (weighted) majority vote among classifiers is a common way to derive the eventual class label for the test instance, but in the case of highly heterogeneous data the majority class in a binary classification problem will seldom be the correct one: rather, it will be highly biased towards the negative class. Therefore, in our implementation the test instance is given the class label of the classifier that assigned it with the highest confidence. details these steps as pseudocode.

Algorithm 1. Message classification using disaster-based ensembles.

Input: Disaster types: $D = \{d_1, d_2, ..., d_M\}$

Instances: $X = \{x_1, x_2, ..., x_{|X|}\}$, each x assigned to $d \in D$.

Class labels: $Y = \{y_1, y_2, ..., y_{|Y|}\}$, $y \in \{-1, +1\}$

/* Training phase */

foreach m *in* M **do**
  train classifier h_m on $\{x|x \in d_m\}$
end

/* Testing phase */

for n = *1 to* $|X|$ **do**
  for m = *1 to* $|D|$ **do**
    obtain class label $\hat{y}_m = h_m(x_n)$ and classifier score s_m
  end
  output the class label $\hat{y}(x_n) = argmax_{(k \in Y)}\{s_m|h_m(x_n) = k\}$
end

Experimental Design

Data

In the experiments that follow, we use the labeled part of the publicly available CrisisLexT26 data set (Olteanu, Vieweg, & Castillo, 2015), which was also studied in a number of previous studies on detection of crisis-related messages in social media, for example, Burel and Alani (2018) and Derczynski et al. (2018). The data set includes tweets on 26 mass disaster events that occurred in 2012 and 2013. The types of emergencies are very diverse and range from terrorist attacks and train derailment to floods and hurricanes (Table 1). The CrisisLexT26 data set was originally created by first retrieving tweets based on a set of search terms relating to specific mass emergencies. The collection can thus be understood to be representative of data that are likely to be found in real-world use cases after elaborated keyword-based filtering. In total, the labeled data set contains 27,933 tweets.

Classification Problems

The proposed methods were evaluated on four different classification problems. These particular classification tasks were chosen because they are all of considerable practical value to emergency responders, and represent different aspects of information that emergency services require to obtain situational understanding (Olteanu et al., 2015). At the same time, these problems differ in terms of the difficulty of classification, each characterized by a different number of categories, balance between categories, and so forth. These problems are (a) *Relatedness*, (b) *Informativeness*, (c) *Topics*, and (d) *Eyewitnesses.* Table 2 provides descriptions and examples of previous research that addressed these classification problems.

Tables 3 and 4 describe class frequencies in the four tasks in the data set. Tables 5 and 6 show examples of

DOI: 10.1002/asi

TABLE 1. Disasters included into CrisisLexT26, their category and the number of hand-labeled tweets.

Disaster	Hazard category	Number of tweets
2012 Colorado wildfires	Wildfire	1,200
2012 Costa Rica earthquake	Earthquake	1,412
2012 Guatemala earthquake	Earthquake	1,050
2012 Italy earthquakes	Earthquake	1,000
2012 Philippines floods	Floods	1,000
2012 Typhoon Pablo	Typhoon	1,000
2012 Venezuela refinery	Explosion	1,000
2013 Alberta floods	Floods	1,000
2013 Australia bushfire	Wildfire	1,199
2013 Bohol earthquake	Earthquake	1,000
2013 Boston bombings	Bombings	1,000
2013 Brazil nightclub fire	Fire	1,000
2013 Colorado floods	Floods	1,000
2013 Glasgow helicopter crash	Crash	1,100
2013 Lac Megantic train crash	Derailment	1,000
2013 LA airport shootings	Shootings	1,032
2013 Manila floods	Floods	1,000
2013 NY train crash	Derailment	1,000
2013 Queensland floods	Floods	1,200
2013 Russia meteor	Meteorite	1,442
2013 Sardinia floods	Floods	1,000
2013 Savar building collapse	Collapse	1,250
2013 Singapore haze	Haze	1,000
2013 Spain train crash	Derailment	1,000
2013 Typhoon Yolanda	Typhoon	1,048
2013 West Texas explosion	Explosion	1,000

positive and negative messages for the four tasks (examples taken from Olteanu et al., 2015).

Preprocessing

We apply a number of preprocessing steps to the data, which are commonly used for Twitter messages before

TABLE 2. Classification tasks.

Task	Description	Previous studies
Relatedness	Separate messages related to a mass emergency from unrelated ones	Li et al. (2015), Burel and Alani (2018)
Informativeness	Identify informative messages (whether the message contributes to better understanding of the crisis situation) as opposed to uninformative ones (refers to the crises but involves sympathy, prayers, and so forth)	Caragea et al. (2016), Derczynski et al. (2018)
Topics	Classify informative messages into six topical categories: Affected individuals, Infrastructure and utilities, Caution and advice, Donations and volunteering, Sympathy and support, Other useful information	Burel and Alani (2018)
Eyewitnesses	Detect eyewitness accounts of mass emergencies (first-hand descriptions of the events)	Imran et al. (2013)

TABLE 3. The sizes of the positive and negative classes in the Relatedness, Informativeness, and Eyewitnesses tasks.

	Positive	Negative	Total
Relatedness	24,581	2,863	27,444
Informativeness	16,849	7,732	24,581
Eyewitnesses	2,193	22,396	24,589

TABLE 4. The sizes of the classes in the Topics task.

	Number of tweets
Affected individuals	4,790
Infrastructure and utilities	1,599
Caution and advice	2,306
Donations and volunteering	2404
Sympathy and support	4,650
Other useful information	7,627
Total	23,376

TABLE 5. Examples of messages in the Relatedness, Informativeness, and Eyewitnesses tasks

	Positive	Negative
Relatedness	RT @NWSBoulder Significant flooding at the Justice Center in #boulderflood	#COstorm you are a funny guy lol
Informativeness	Flash floods wash away homes, kill at least one near Boulder via @NBCnews	Pray for Boulder, Colorado #boulderflood
Eyewitnesses	Outside sounds like it is going to shatter my bedroom windows any sec now #bigwet #qld	RT @RedCrossAU: Everyone affected by #qldfloods, let people know you are safe: http://t.co/..

TABLE 6. Examples of messages of the classes in the Topics task.

	Examples
Affected individuals	Colorado fire displaces hundreds; 1 person missing: Firefighters in Colorado and New Mexico are battling wind-fu... http://t.co/R6OQwpix
Infrastructure and utilities	The High Park fire west of Fort Collins, #CO has consumed 36,930 acres so far, is 0% contained and continues to grow. #NWS #cowx #cofire
Caution and advice	RT LarimerSheriff: #HighParkFire evacuation orders issued for Pingree Park area. 25 notifications sent http://t.co/oSmxBfqJ
Donations and volunteering	#offer @nocok9cop We can take in a couple or small family + pets at our house for evacuees of #highparkfire - #loc live in Windsor.
Sympathy and support	RT @hannadianee: Hope everyone's ok #prayforboston
Other useful information	FEMA has authorized the use of federal funds to help with firefighting costs for the #HighParkFire. http://t.co/whxlpPEP

DOI: 10.1002/asi

performing text classification. Before linguistic processing of the message, the text was normalized in the following way: We removed mentions, URLs, sequences of hashtags at the start and end of the message, and word tokens consisting of digits were replaced with a unique tag. The normalized text was tagged for parts-of-speech using Pattern (De Smedt & Daelemans, 2012).

Classification Methods

To train classifiers, we experiment with the following algorithms that have been previously often used for short-message classification (for example, Ashktorab et al., 2014; Caragea et al., 2016; Li et al., 2015):

K-nearest neighbor (kNN). The kNN algorithm classifies a test instance by first identifying its k-NNs among the training instances according to some similarity measure and then assigning it to the class that has the majority in the set of nearest neighbors. We set k to be equal to 5, via fine-tuning experiments.

Multinomial Naïve Bayes (MNB). MNB implements the Naïve Bayes algorithm for multinomially distributed data. It has been shown to perform better than simple Naïve Bayes, especially at larger vocabulary sizes (McCallum & Nigam, 1998).

Decision tree (DT). A DT classifier is an inductive rule algorithm that during training builds a tree, in which nodes correspond to features, branches departing from them are determined by the weight of the feature in the data (for example, Information Gain), and leafs are class labels. During testing, a DT classifier classifies a test document by traversing the tree along the paths determined by its features, until a leaf node is reached.

Maximum entropy (MaxEnt). The MaxEnt (a.k.a. logistical regression) algorithm is a probabilistic classification method based on the Principle of Maximum Entropy: from all the models that fit the training data, it selects the one that has the largest entropy. Unlike the Naïve Bayes classifier, MaxEnt does not assume that the features are conditionally independent of one another, and so often leads to better results for text classification, where features are natural language words with a high degree of interdependence.

Support vector machines (SVM)

SVM is a function-based classifier built upon the concept of decision planes that define class boundaries. In our experiment, we use the linear SVM with $C = 1.0$. SVM has been known to be among the superior learning methods for text classification. We use the scikit-learn implementations of these algorithms.[1,2]

[1] USAF DISTRIBUTION STATEMENT A. Approved for public release: distribution is unlimited. 88ABW Cleared 1/30/2016; 88ABW-2016-6141.

[2] http://scikit-learn.org/stable/ (retrieved on March 21, 2018).

Evaluation Metrics

The quality of classification was measured in terms of the traditional measures of precision, recall, and F-measure. For a given category, precision is a measure of accuracy and is the percent of correct predictions out of all predictions for that category. Recall is a measure of sensitivity and is the percent of correct predictions out of all samples in that category. Because in the Relatedness, Informativeness, and Eyewitnesses tasks, the problem is a binary classification, and our main interest is in the positive category, we report these measures only for the positive category. For the Topics tasks, the reported measures are macro-averages over all the six categories. The F-measure is a geometric mean of the precision and recall, which aims to discourage big differences in precision and recall of a particular classifier. In calculating an F-measure, we give an equal weight to precision and recall.

Cross-Validation Scenarios

We conducted experiments with three scenarios reflecting possible application scenarios of a system for detecting disaster-related messages.

Scenario 1. A classifier was trained and tested on data representing the same disaster. This scenario assumes that within a practical application, a classifier is trained on messages that refer to a particular disaster event and that the messages are collected using very detailed and precise keyword searches and manually labeled in real time, possibly via crowd-sourcing to a team of volunteers or paid workers (for example, Imran et al., 2014). The scenario corresponds to the closest fit between the training data and the test data. The data in each of the 26 disaster sets was randomly split into 10 parts, and each classifier was evaluated using the usual 10-fold cross-validation technique. The eventual performance was measured by averaging precision, recall and F rates of the 26 classifiers. Examples of previous work that evaluated their classification models under this scenario includes Caragea et al. (2011), Musaev et al. (2014), and Caragea et al. (2016).

Scenario 2. In the second use case scenario, the entire data set was used to train and test a single classifier, which was evaluated using 10-fold cross-validation. Examples in the data set were distributed into training and test parts randomly, which ensured that data on the same crisis was present in both training and test data. This scenario is more challenging than the first, as the classifier needs to generalize over data on multiple disasters; at the same time, because the test data are drawn from the same (multiple) distributions as the training data, this classification problem is not affected by data heterogeneity. This application scenario was assumed in previous studies by Ashktorab et al. (2014), Burel and Alani (2018), and Derczynski et al. (2018).

Scenario 3. The third scenario reflects the use case where messages that need to be classified represented disasters,

DOI: 10.1002/asi

whose types are not known in advance. The train–test split was done in a way such that the test data contained tweets only on those crises that were not included into the training data, that is, simulating the conditions when a disaster needs to be detected before any manually labeled data relating to it are available. Specifically, at each split data on 23 crises were used for training and data on the three remaining crises were used for testing. The reported performance scores are averages over nine such splits. To our knowledge, a similar scenario was previously evaluated only in studies by Verma et al. (2011) and Imran et al. (2013), but whereas these articles trained a model on one disaster event and tested on another, Scenario 3 in this article refers to training on one set of multiple events and testing on another set of multiple events, that is, a more realistic and harder application scenario.

Experiments

Effect of Data Heterogeneity

In the first experiment, we compared the difficulty of the classification problems under the three scenarios, specifically aiming to determine how much degradation the performance of a classifier is likely to suffer when deployed under Scenario 3 vs. Scenarios 1 and 2. We evaluated the five base learning methods—kNN, MNB, DT, SVM, and MaxEnt on each of the scenarios. Figure 1 presents the results of these runs.

For the Relatedness task, the learning methods perform similarly under Scenarios 1 and 2, with the exception of SVM. F-measures for Scenario 3 are lower than the other two, although the performance drop is never less than 2%. The Relatedness task appears to be an easy problem, with all the classifiers achieving uniformly high levels of F-measure.

In the Informativeness task, Scenarios 1 and 2 also show similar F-measure rates, although here the results for Scenario 2 are somewhat better, by 2–3%. Scenario 3 is behind Scenarios 1 and 2, but only insignificantly, by 4% to 5%. As with the Relatedness task, the difference between the learners is not very high, with each of them achieving an F-measure above 80 for all the scenarios.

The Topics task proved harder than the preceding two. The learners obtain F-measures mostly between 40% and 60%. Under Scenario 3, SVM and MaxEnt show comparable results (F of 50.7% and 53.3%, correspondingly), outperforming the other three learners by 7% to 16%. Interestingly, Scenario 2 is a simpler problem than Scenario 1, which suggests that for this

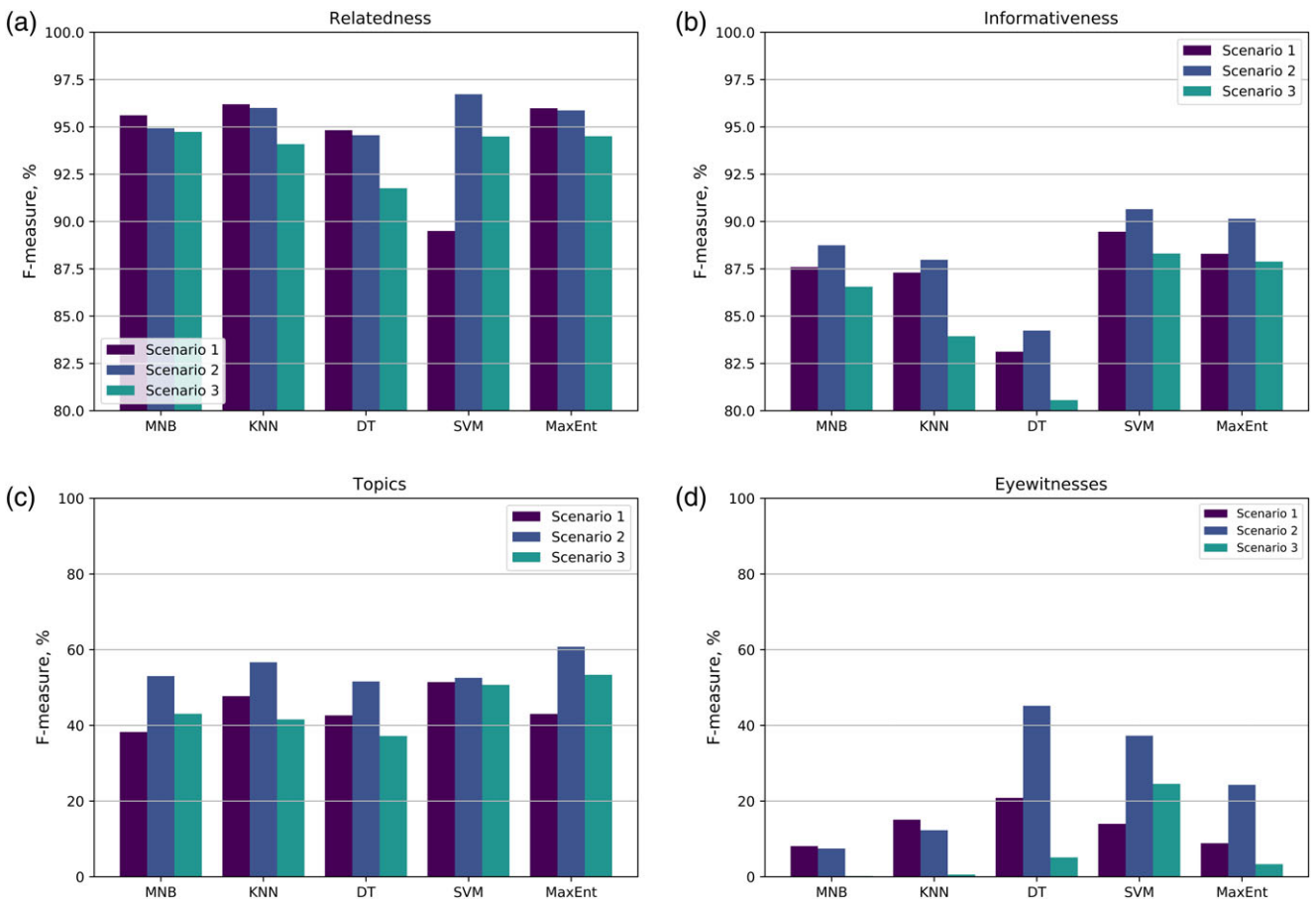

FIG. 1. Comparison of F-measures achieved by the base learning methods on the Relatedness, Informativeness, Topics, and Eyewitnesses tasks. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

Topics task, the scarcity of training data available for a specific disaster event outweighs the greater match between the training and test set. As before, Scenario 3 yields worse results than Scenario 2, across all the learners; this time the drop is much greater, up to 14%.

The Eyewitnesses task is the hardest, with none of the learners reaching the F-measure of over 50%, under none of the scenarios. DT, SVM, and MaxEnt fare better than Naïve Bayes and kNN, and Scenario 2 is a much easier problem than the other two. Under Scenario 3, the best performer is SVM, with F-measure of 24.5%, but across all learners the performance is noticeably worse compared to Scenarios 1 and 2.

Thus, Scenario 3 leads to poor efficacy for the Topics and the Eyewitnesses tasks, apparently due to discrepancies between the training and test sets. To verify this, we measured the difference between feature distributions of training and test data under Scenario 2 vs. Scenario 3. Obtaining probabilistic feature distributions via Maximum Likelihood Estimation, we measured the Jensen-Shannon divergence, a variant of Kullback–Leibler divergence that ranges between 0 and 2, between the training and test set in each train–test split. We found that the mean Jenson–Shannon divergence in Scenario 2 is 0.01, whereas in Scenario 3 it is 0.07; the difference is significant based on an independent samples *t*-test ($p < .001$), confirming that there is indeed a much greater difference between the training and test data under Scenario 3. These results therefore are consistent with the findings by Verma et al. (2011) and Imran et al. (2013) that one can expect a significant loss in classifier accuracy when a model is trained on one disaster type, but tested on another.

Ensemble Classification

We next examined whether or not ensemble methods can improve on the performance of the base learners under Scenario 3. The experiments included AdaBoost, GBC, and Random Forests, a DT classifier, which is used as a base learner in these methods, as well as three disaster-based (DB) ensembles, where base learners are DT, SVM, and MaxEnt. These results are presented in Figure 2.

On the Relatedness task, all the ensemble methods perform very similarly, with all of them improving on DT in terms of recall, which also leads to a high F-measure.

On Informativeness, all the ensembles outperform DT in terms of recall, which also yields a better F-measure, by 3–5%. It is worth noting that disaster-based ensembles obtain better recall rates than Random Forest, AdaBoost, GBC, and DT reaching, but also produced lower precision than these methods. The results achieved by the classifiers are comparable to the F-scores achieved in previous studies on the informativeness problem: for example, 0.91 in Derczynski et al. (2018).

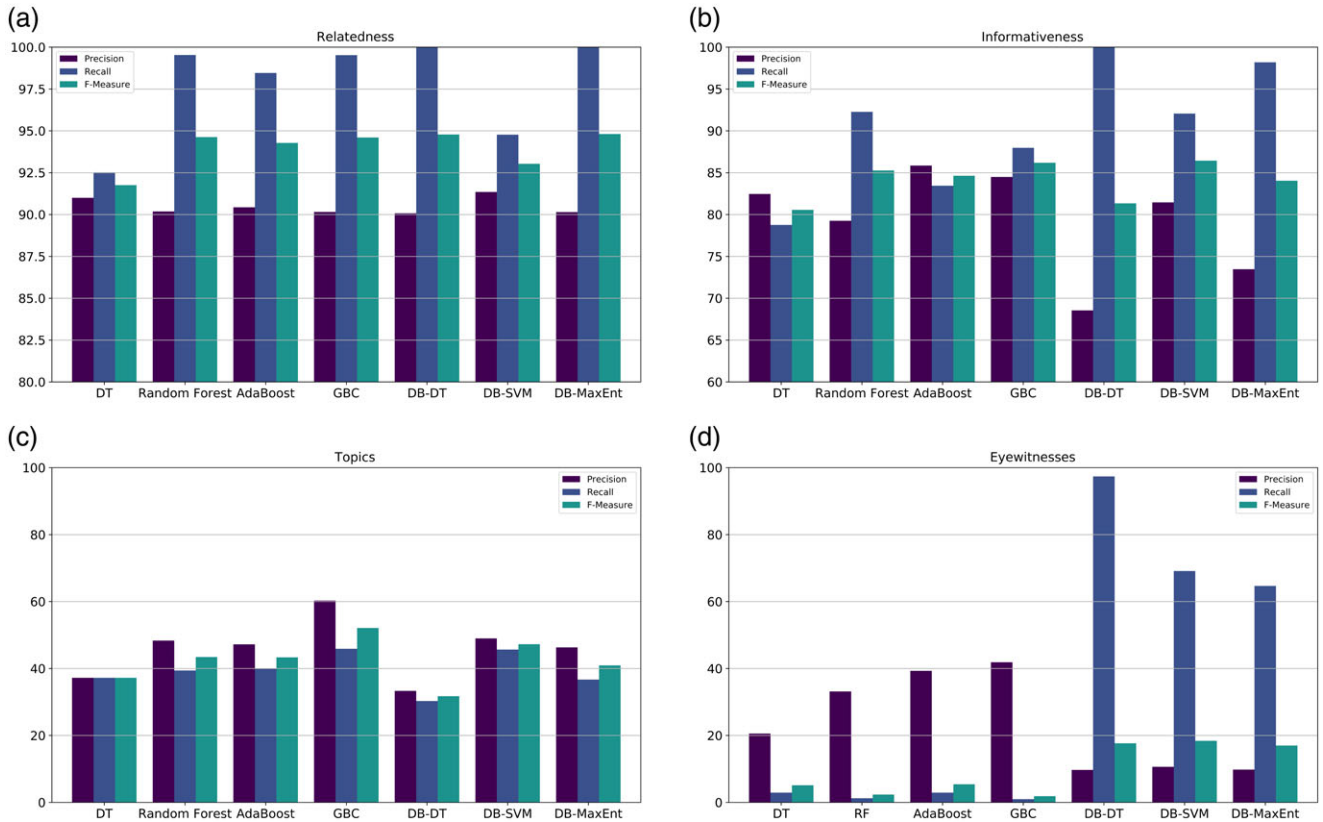

FIG. 2. Performance of ensemble classifiers for the four evaluation tasks. [Color figure can be viewed at wileyonlinelibrary.com]

DOI: 10.1002/asi

On Topics, all the ensembles outperform DT (with the exception of DB-DT) in terms of both precision and recall. The best results are achieved by GBC, which improves precision by 23% and recall by 8% on DT. Disaster-based SVM and MaxEnt ensembles fare similarly to Random Forests and AdaBoost, the precision differences being no more than 2% and recall no more than 9%. The F-score of GBC (0.53) is somewhat lower than the F-score shown by convolutional neural networks (0.61), the best-performing classifier in the study by Burel and Alani (2018), evaluated on the same data set, although our evaluation scenario is more difficult that the one used in Burel and Alani (2018), whose experimental design is similar to Scenario 2 used in this study.

On Eyewitnesses, disaster-based ensembles showed very high recall rates compared to Random Forests, AdaBoost, and GBC, but also lower precision, with the F-measure still superior to those of the other three ensembles.

Overall, we find that ensemble classifiers do tend to perform better than base classifiers under Scenario 3. The proposed disaster-based ensembles generally perform on a par with the popular Random Forests, AdaBoost, and GBC ensembles; the differences between the two groups are significant only on the Eyewitnesses task, where the former produce higher recall, while for the latter, higher precision.

Discussion of Results

Heterogeneity in both training and test data is known to present a major problem for machine-learning algorithms. Our first set of experiments confirmed that this is indeed the case, with short messages relating to multiple and highly diverse disaster events: the accuracy of five different base classifiers was found to degrade significantly when switching from Scenario 1 (training and testing on data about the same disaster event) and Scenario 2 (training and testing on the same set of events) to Scenario 3 (training on some event types, while testing on others). However, we find that data heterogeneity does not affect the relatively simple tasks of finding messages that are disaster-related or informative, that is, contributing to situational awareness. Its adverse effects are significant when classifying messages into semantic categories and determining eyewitness accounts among them. We also find that under Scenario 3, compared to Scenario 2, there is indeed a greater divergence between the training and test data sets in terms of feature distributions, indicating that this must be the reason for the accuracy drop.

Subsequent experiments were concerned specifically with Scenario 3, as this is the most likely practical use case, that is, when automatic detection and classification of relevant messages are required without any prior knowledge of the type of the disaster they represent. Our purpose here was to investigate ensemble learning methods as a means to improve on the classification accuracy achieved by base classifiers under this usage scenario.

The results of the experiments with ensemble methods show that, overall, they do tend to perform better under Scenario 3 than base classifiers, either only in terms of recall, or both precision and recall. The newly proposed disaster-based ensembles generally perform on a par with the popular Random Forests, AdaBoost, and GBC ensembles; the differences between them are significant only on the Eyewitnesses task, where the former ensembles produce higher recall, while the latter, higher precision.

Thus, we can offer the following general recommendations for future practical applications in use cases similar to Scenario 3. Data heterogeneity does not cause significant problems for base classifiers under the relatively easy Relatedness and Informativeness tasks, where they achieve high levels of both precision and recall and where more sophisticated techniques do not yield any benefits. But for the Topics and Eyewitnesses tasks, the two harder classification problems, all ensemble methods produce uniformly better results than base classifiers, particularly in terms of recall. If information about the types of disasters is available in the training data, the new proposed ensemble method that takes advantage of this information tends to fare better than traditional methods like AdaBoost, Random Forest, and GBC in terms of recall, but not precision.

A Use-Case Study

In this section, we test the ability of ensemble learners that proved to be best-performing in previous experiments to generalize to real-world data. Since it is practically impossible to measure the recall on real-world data (it is impossible to know all the messages on Twitter that belong to a category), we were interested in determining precision of the methods.

Data collection and classification. The real-world data used in this experiment consisted of around 2.4 million tweets collected using 24 single-word queries that refer to different kinds of disasters via Twitter Search API. The tweets obtained using generic queries were assigned labels in the following manner. Three classifiers were trained on the CrisisLex data set. Based on the results of the previous experiments, we used the MaxEnt classifiers for the Relatedness and Informativeness steps, where it proved to produce the highest accuracy. For the Eyewitnesses classifications, we used the GBC, which demonstrated the highest precision on this task. First, the Relatedness classifier was used to detect messages relevant to a disaster. Then the Informativeness classifier was used to identify informative messages among those that were classified as related in the previous step. Finally, the Eyewitnesses classifier was run on the informative messages to detect eyewitness accounts among them.

Human judgments. We selected 150 messages that the Eyewitness classifiers labeled as positive examples with the greatest confidence scores. We then asked two human judges to evaluate these messages: the judges were instructed to mark each message as (a) being informative or not, and (b) as containing an eyewitness account of an emergency

DOI: 10.1002/asi

TABLE 7. Tweets from the real-world data set judged to be informative and containing eyewitness reports by human judges.

Informative	Eyewitness reports
#SNHR #SyriaMrs. Fawziyeh Al Nawfal from Hasakah died in unknown source landmine explosion in Makhroom area in Hasakah, Mar 7	Legit cannot step outside of my house since there is a giant storm and the warmest thing I have is a bathing suit cover up....
News: Deal man Christian Sloan died in tragic waterfall accident in #Vietnam - Kent Online https://t.co/Lg1828y6U9	Will be filling my day with Overwatch. Snow storm means husband drove my vehicle to work.. and I am stuck in the home.: 3
Father died in crash on way home after meeting his baby https://t.co/WDsCzJFSLH	That's clearly the worst storm since the beginning of the winter, all the road are closed, that really suck, my training on the ice is stuck

situation or not. In the instructions we used definitions for "informativeness" and "eyewitness reports" similar to those used by Olteanu et al. (2015) in constructing the CrisisLex data set. Table 7 shows three randomly selected tweets that were judged to be informative and eyewitness reports by both judges.

The Cohen's κ statistics for the agreement between the judges was 0.48 for the Informativeness judgments and 0.60 for the Eyewitnesses judgments; both figures within the range that is normally taken to indicate moderate agreement (Landis & Koch, 1977). The κ values we obtain are similar to those reported in Olteanu et al. (2015), where they find that the agreement between individual annotators on labeling the source of the disaster-related information (which includes eyewitness accounts) is between 0.57 and 0.63. This level of agreement can be taken as an indication of the upper bound on the performance of the classifiers that can be expected on real-world data.

Results and error analysis. Table 8 shows the precision of the two classifiers determined relative to the two judge's labels. The precision rates obtained in this experiment are lower, but generally consistent with those obtained in experiments with the CrisisLex data set, where the MaxEnt classifier reached 86.2% for the Informativeness task, and the GBC achieved 41.8% for Eyewitnesses.

To understand the reasons for errors made by the classifiers, we looked at cases where both judges believed the classifiers assigned the wrong label and identified common error types: (a) news reports on accidents that are irrelevant to any rescue operations, (b) errors due to ambiguous words, (c) disaster events that took place far in the past, (d) fictional events (movies, song lyrics, and so forth), (e) general chatter. The percentage and examples of these error types are shown in Table 9.

TABLE 8. Precision of the Informativeness and Eyewitnesses classifiers on the real-world data.

	Informativeness	Eyewitnesses
Judge 1	73.37%	66.8%
Judge 2	81.1%	62.9%

TABLE 9. Tweets with the most common error types in the real-world data set.

Error type	Percent	Example
News	67.4	RT @CFRAOttawa: Sources confirm Ottawa firefighter Shawn Mathieson died today in a snowmobile crash. He had two children. #ottnews
Word ambiguity	16.3	To the chainsaw massacre going on to the trees outside my house... let us start at like 10 from now on. 6 o'clock a.m. is out of control
Past events	6.1	RT @clandro: #TodayinHistoryMarch 5,1963: American country singer Patsy Cline died in a plane crash. https://t.co/gR8OFhQmiR
General chat	6.1	Saw #AlexanderSkarsgard was credited as Adam in #Zoolander2 I was like um no Meekus died in a freak gasoline fight accident
Fictional events	4.0	AU // What if Arizona died in the explosion? #GreysAnatomy #CallieTorres #Arizon... (Vine by @HeelyQueen) https://t.co/da6qwBgOO1

The most common error type, the news reports, account for around 66% of all errors and seem very difficult to distinguish from informative messages: a news item does not directly state if rescue operations are ongoing or are already over. They are also difficult to distinguish from eyewitness reports, as their content and style are very similar. Regarding other error types, classification of messages involving ambiguous words can potentially be improved using extra training data and/or additional Natural Language Processing (NLP) techniques, such as word sense disambiguation. Other types of errors may require special classifiers that would recognize the time references in messages, and whether a message describes a fictional event.

Conclusion

In this article, we explored text classification methods that would be suitable for application in practical, real-world scenarios, where the monitoring system is tasked with identifying reports of potential emergency situations without prior knowledge of either specific events or their type. Such use case scenarios are characterized by high heterogeneity of the data, which causes significant performance degradation for text classifiers.

The contributions of the article can be summarized as follows. We provide a study of the effect that data heterogeneity has on nonensemble classifiers in the context of detecting disaster-related messages. We demonstrate that training classifiers on some types of disasters, while testing on other ones, leads to a significant drop, both in precision and recall, on the four classification problems relevant to disaster management. To deal with data heterogeneity, we

DOI: 10.1002/asi

introduced a new ensemble learning method that makes use of information about disaster types available in training data. Our experiments show that this method clearly outperforms base classifiers and performs on a par with several other popular ensemble classifiers (AdaBoost, Random Forests, Gradient Boosting). Finally, in a use case study, we verified the ability of our proposed methods to handle the massive diversity of real-world social media data, for the first time obtaining results indicating likely performance levels that can be achieved in practical real-world applications.

There is clearly much work to be done. The goal is far more important than the mere correct classifications of data for assessing the scope of situational awareness problems. The ultimate objective is to create a reliable tool that allows first responders to leverage social media to ensure the safety of the public at large. The testimony of the value of such a tool occurs when those who utilize this research in their work areas are able to improve the success rates of their recovery operations in times of real crises.

Acknowledgments

Professor Jane M. Binner and Dr. Viktor Pekar gratefully acknowledge the financial support from the University of Birmingham's Maxwell Fry Global Finance research fund, courtesy of Tokai Bank.

References

Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining twitter to inform disaster response. In *ISCRAM*.

Ballings, M. & den Poel, D.V. (2015). CRM in social media: Predicting increases in Facebook usage frequency. European Journal of Operational Research, 244(1), 248–260.

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. In Fifth International AAAI Conference on Weblogs and Social Media.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Burel, G. & Alani, H. (2018). Crisis event extraction service (CREES) — automatic detection and classification of crisis-related content on social media. In 15th International Conference on Information Systems for Crisis Response and Management. Retrieved from http://oro.open.ac.uk/55139/

Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P. ... Yen, J. (2011). Classifying Text messages for the haiti earthquake. In ISCRAM.

Caragea, C., Silvescu, A., & Tapia, A.H. (2016). Identifying informative messages in disasters using convolutional neural networks. In ISCRAM.

Choo, C.W., & Nadarajah, I. (2014). Early warning information seeking in the 2009 Victorian bushfires. Journal of the Association for Information Science and Technology, 65(1), 84–97.

Chua, A.Y. (2007). A tale of two hurricanes: Comparing Katrina and Rita through a knowledge management perspective. Journal of the American Society for Information Science and Technology, 58(10), 1518–1528.

De Smedt, T. & Daelemans, W. (2012). Pattern for python. Journal of Machine Learning Research, 13(1), 2063–2067.

Derczynski, L., Meesters, K., Bontcheva, K., & Maynard, D. (2018). Helping crisis responders find the informative needle in the tweet haystack. In ISCRAM 2018 Conference Proceedings 15th International Conference on Information Systems for Crisis Response and Management. Retrieved from http://oro.open.ac.uk/55139/

Dietterich, T. G. (2000). Ensemble methods in machine learning. In Proceedings of the First International Workshop on Multiple Classifier Systems (MCS'00) (pp. 1–15). London: Springer.

Duan, L.L., Clancy, J.P., & Szczesniak, R.D. (2016). Bayesian ensemble trees (bet) for clustering and prediction in heterogeneous data. Journal of Computational and Graphical Statistics, 25(3), 748–761.

Fan, C., Mostafavi, A., Gupta, A., & Zhang, C. (2018). A system analytics framework for detecting infrastructure-related topics in disasters using social sensing. In I.F.C. Smith & B. Domer (Eds.), Advanced computing strategies for engineering (pp. 74–91). Cham, Switzerland: Springer International Publishing.

Freund, Y. & Schapire, R.E. (1996). Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on International Conference on Machine Learning (ICML'96) (pp. 148–156). Bari, Italy: Morgan Kaufmann Publishers.

Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

Hagen, M., Potthast, M., Büchner, M., & Stein, B. (2015). Twitter sentiment detection via ensemble classification using averaged confidence scores. In Advances in Information Retrieval - 37th European Conference on IR Research, ECIR (pp. 741–754).

Heverin, T. & Zach, L. (2012). Use of microblogging for collective sensemaking during violent crises: A study of three campus shootings. Journal of the American Society for Information Science and Technology, 63(1), 34–47.

Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). AIDR: Artificial intelligence for disaster response. In Proceedings of the 23rd International Conference on World Wide Web (WWW'14) (pp. 159–162). Companion Seoul, Korea: ACM.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Extracting information nuggets from disaster-related messages in social media. In ISCRAM.

Imran, M., Mitra, P., & Srivastava, J. (2016). Cross-language domain adaptation for classifying crisis-related short messages. In 13th Proceedings of the International Conference on Information Systems for Crisis Response and Management.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C. ... Tapia, A.H. (2015). Twitter mining for disasters response: A domain adaptation approach. In ISCRAM.

Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., & Miller, R.C. (2011). Twitinfo: Aggregating and visualizing microblogs for event exploration. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'11) (pp. 227–236). Vancouver, BC, Canada: ACM.

McCallum, A. & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In Learning for Text Categorization: Papers from the 1998 AAAI Workshop (pp. 41–48).

Musaev, A., Wang, D., Cho, C.A., & Pu, C. (2014). Landslide detection service based on composition of physical and social information services. In 2014 I.E. International Conference on Web Services (pp. 97–104).

Nguyen, D., Mannai, K.A.A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional. Neural Networks, 1, 632–635. Retrieved from https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15655

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'15) (pp. 994–1009). Vancouver, BC, Canada: ACM.

Schmidt, V.A. & Binner, J.M. (2015). A semi-automated display for geotagged text. In J. Preston, J.M. Binner, L. Branicki, T. Galla, N. Jones, J. King, et al. (Eds.), City evacuations: An interdisciplinary approach (pp. 107–116). Berlin, Heidelberg: Springer.

Tuarob, S., Tucker, C.S., Salathe, M., & Ram, N. (2014). An ensemble heterogeneous classification methodology for discovering health-related

knowledge in social media messages. Journal of Biomedical Informatics, 49, 255–268.

Verma, S., Vieweg, S., Corvey, W., 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 L.A. Adamic, R.A. Baeza-Yates, & S. Counts (Eds.), ICWSM. Palo Alto, CA: AAAI Press.

Yates, D. & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. International Journal of Information Management, 31(1), 6–13.

DOI: 10.1002/asi