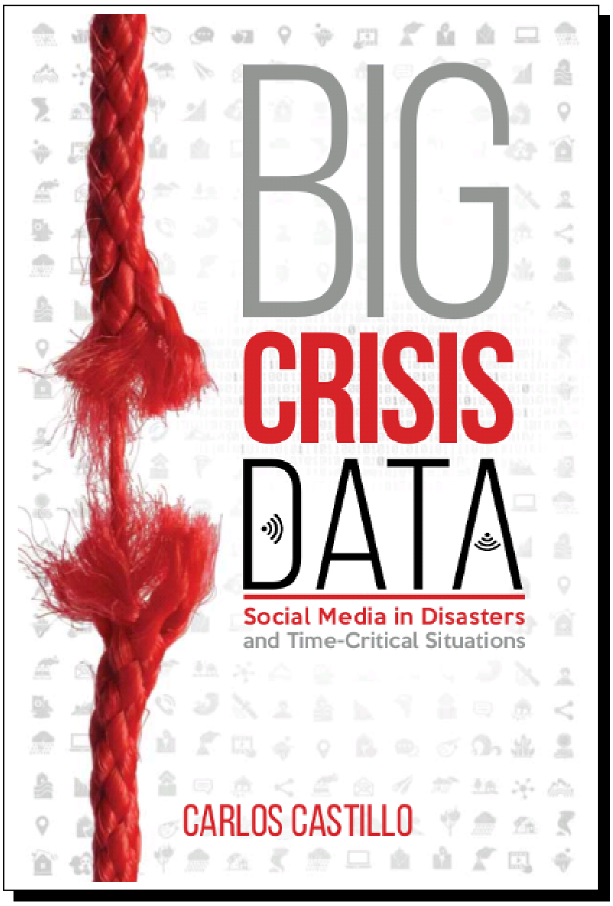
Big Crisis Data

Social Media in Disasters and Time Critical Situations

by Carlos Castillo



FREE PREVIEW

CHAPTER 4. “VARIETY”

Cite as: Carlos Castillo: Big Crisis Data. Cambridge University Press, 2016

BibTex:

```
@book{castillo2016big,
      title="Big Crisis Data",
      author="Castillo, Carlos",
      publisher="Cambridge University Press"
      year=2016
}
```

Contents of Chapter 4

4

Variety: Classification and Clustering

In 2005 the Inter-Agency Standing Committee (IASC), a permanent forum including agencies from the United Nations (UN) and agencies not belonging to the UN (such as the Red Cross), introduced a number of reforms designed to improve humanitarian response. A visible reform was the establishment of the *Cluster System*,[1] which organizes large-scale multiagency humanitarian response into eleven areas of action, each one with its own responsibilities: health, protection, food security, emergency telecommunication, early recovery, education, sanitation, water and hygiene, logistics, nutrition, emergency shelter, and camp management and coordination. The Cluster System is not without critics, but it serves to structure response and it is liked by national governments because it introduces a single focal point which is accountable for a specific response area.[2]

This chapter describes methods for automatic text categorization, which allow us to make sense of heterogeneous, *varied* messages by sorting them into categories. In the same way in which coordination among humanitarian agencies is facilitated by abstracting from specific response actions to response areas, coping with typical crisis collections from social media, involving millions of messages, is made easier by abstracting from the particular (a specific message) to the general (a class of messages).

There are two broad families of classification methods: supervised and unsupervised. In *supervised classification*, we first manually classify a set of items (messages in this case) into categories using human annotators, and then use these example items to automatically learn a model for classifying new, unseen items into the same categories. In *unsupervised classification* (or *clustering*), we do not provide any example item classified a priori, but instead allow a method to discover groups of related items based on their similarity.

We begin with a description of the main information categories found in social media and short text messages during crises (§4.1). Next, we introduce supervised (§4.2) and unsupervised classification methods (§4.3).

4.1 Content Categories

The first question when categorizing content is how to determine which information categories to use. There are many factors that drive the design of these categories. The first and most important are the *information needs* of the users for which the categorization is done, which may include emergency managers, humanitarian relief workers, policy makers, analysts, and/or the public. Different audiences may have different information needs.

When performing automatic classification, it is also important to take into account the *capabilities of the system* that will be used to categorize the messages. Even with a good data representation and with a state-of-the-art learning method, current computational methods for text classification are more accurate in distinguishing among information categories that have clearly distinct vocabularies, than in distinguishing those having many characteristic terms in common. Other tasks may be too hard for computers, such as determining if a message is sarcastic or funny, literal or metaphoric.

When performing manual classifications, the *capabilities of humans* constraint our choice of categories in the sense that we cannot deal with a number of categories that is too large. Using nonexpert annotators imposes another constrain: categories that are too fine-grained or involve subtle distinctions cannot be reliably annotated

[1] UN OCHA: "The Cluster System" http://www.unocha.org/what-we-do/coordination-tools/ cluster-coordination.
[2] UN OCHA Annual Report 2006. http://www.unocha.org/annualreport/2006/html/part1 _humanitarian.html.

by nonexperts. Particularly in the case of volunteer annotations, it is best to have few, easily understandable categories (more on this in Chapter 7).

Another aspect that is relevant to consider for some applications is the *availability of information*, that is, what are the categories of information the public are actually sharing and discussing in social media. All other things equal, it might be more useful to use a typology that is more fine-grained in the information categories that are more abundant and more coarse-grained in the ones that are less abundant.

Existing typologies. A number of typologies have been proposed to categorize social media messages during crises. With the exception of systematic work in ontologies for crisis information, such as the ones presented in Section 3.7, most typologies are created based on observations from a small number of crisis situations. Existing typologies for crisis-related social media described in the literature cover many dimensions for categorizing content:

(i) By factual, subjective, or emotional content: to separate between messages conveying facts (or combinations of facts and opinions), and messages conveying opinions, or emotions, such as expressions of sympathy.

(ii) By information provided: to extract particular categories of information that are useful to experts or the public for various purposes.

(iii) By information source: to select messages posted by specific groups of users, for instance, messages by non-governmental organizations or messages from official government sources.

(iv) By credibility: to filter out messages that are unlikely to be considered credible, or whose authenticity can be questioned (more on this in Chapter 8).

(v) By time: to filter messages that refer to different stages of an event (in the sense of the natural progression of a disaster), when temporal boundaries for the event are unclear.

(vi) By location: for instance, to separate eyewitness accounts from messages posted by people away from the scene.

(vii) By embedded links: some messages may point to other material, including news articles, photos, videos, or live video feeds.

(viii) By high-level environment: a taxonomy by Mileti (1999) that divides information about a disaster into that pertaining the physical environment, the built environment, and the social environment.

Table 4.1 summarizes some of these dimensions and references previous work in which they have been mentioned or described.

Different types of crisis generate different types of messages in social media. In a transversal analysis involving 26 crisis situations, Olteanu et al. (2015) found that sometimes the information type that was the most prevalent in one disaster may be present only in very small amounts in another. This variability can be traced back to several factors, including whether the events are instantaneous or progressive, whether their effects are diffuse or focalized, and whether the causes are natural hazards or human action. We come back to these and other factors that affect content production in social media in Section 9.3.

The distribution of messages into information categories is also affected by time, as expected given the particular characteristics of different stages of a disaster (Petak, 1985; Fischer, 1998). For instance, in a meteorological emergency such as a tornado, the wind speed may be the initial focus of the conversation, while more detailed/useful information may arrive later (Smith et al., 2015). Olteanu et al. (2015) observed a general pattern of progression of information in disasters, with messages of caution and advice arriving first, followed by sympathy and support, followed at the end of the first day or on the second day by messages describing infrastructure damage and affected individuals, followed by useful messages covering various topics, followed from about the third day onward by messages regarding donations and volunteering. Parsons et al. (2015) also observed a general progression of themes following known stages in disaster life cycles.

4.2 Supervised Classification

There are three main scenarios for classification: binary, multiclass, and multilabel. *Binary classification* refers to a categorization into two disjoint classes: for instance, a message can be related to a certain crisis situation, or not. *Multiclass classification* refers to categorization into a series of disjoint classes: for instance a message during a tornado can be about providing advice, soliciting donations, reporting the weather conditions, or other categories.

Table 4.1. *Classification of various dimensions of content posted on social media during high-impact events, including their description and references to related work.*

By factual, subjective, or emotional content	
Factual information	*(Examples under "By information provided")*
Opinions	opinions, criticism (e.g., criticism of government response)
Sympathy	condolences, sympathy (Kumar et al., 2013); concerns and condolences (Acar and Muraki, 2011), support (Hughes et al., 2014); thanks and gratitude, support (Bruns et al., 2012; Shaw et al., 2013); gratitude, prayers (Olteanu et al., 2014); emotional support (Taylor et al., 2012); emotion-related (Qu et al., 2011)
Antipathy	*schadenfreude*, animosity against victims (e.g., because of a long-standing conflict among countries) (Imran et al., 2015)
Jokes	jokes, trolling (Metaxas and Mustafaraj, 2013); humor (Leavitt and Clark, 2014); humor or irrelevant/spam (Sreenivasan et al., 2011)
By information provided	
Caution and advice	caution and advice (Imran et al., 2013); warnings (Acar and Muraki, 2011); advice, warnings, preparation (Olteanu et al., 2014); warning, advice, caution, preparation (Vieweg et al., 2010); tips (Leavitt and Clark, 2014); safety, preparation, status, protocol (Hughes et al., 2014); preparedness (Wukich and Mergel, 28/08/2014); advice (Bruns, 2014); advice and instructions (Shaw et al., 2013); predicting or forecasting, instructions to handle certain situations (Sreenivasan et al., 2011); safety (St. Denis et al., 2014)
Affected people	medical emergency, people trapped, person news (Caragea et al., 2011); casualties, people missing, found or seen (Imran et al., 2013); self reports (Acar and Muraki, 2011); fatality, injury, missing (Neubig et al., 2011; Vieweg, 2012); looking for missing people (Qu et al., 2011)
Infrastructure and utilities	infrastructure damage (Imran et al., 2013); collapsed structure (Caragea et al., 2011); built environment (Vieweg, 2012); damage, closures and services (Hughes et al., 2014); services (St. Denis et al., 2014); collapsed structure, water shortage/sanitation, hospital/clinic services (Caragea et al., 2011); road closures and traffic conditions (Truelove et al., 2014)
Needs and donations	donation of money, goods, services (Imran et al., 2013); food/water shortage/distribution (Caragea et al., 2011); donations or volunteering (Olteanu et al., 2014); help requests, relief coordination (Qu et al., 2011); relief, donations, resources (Hughes et al., 2014); help and fund-raising (Bruns, 2014); volunteer information (Vieweg et al., 2010); help requests (Acar and Muraki, 2011; Neubig et al., 2011); requests and offers of donations (Purohit et al., 2014a,b)
Nonhuman animals	animal management (Vieweg et al., 2010), lost and found pets (White et al., 2014; Barrenechea et al., 2015); animal evacuation (White and Palen, 2015)
Weather and status updates	weather updates (Vieweg, 2012); status (St. Denis et al., 2014); smoke, ash (Truelove et al., 2014)
Other useful information	hospital/clinic service, water sanitation (Caragea et al., 2011); reports about environment (Acar and Muraki, 2011); consequences (Olteanu et al., 2014)

(*cont.*)

Multilabel classification refers to categorization into classes that do not need to be disjoint: for instance, a message can be simultaneously about donations of food and clothes – performing automatic multilabel classification is also referred to as *tagging*, in the understanding that a message can have more than one tag.

There are many ways of performing automatic classification of messages. A straightforward, but ineffective approach, is to use keyword-based rules to separate messages into categories. For instance, a message containing the word "shelter" or "camp" can be associated to the category "emergency shelter." This may work for certain information categories that have a small, well-defined, unambiguous set of terms that are highly discriminative, but in general they are ineffective for categories that lack those terms (Melville et al., 2013).

A more robust approach is to use statistical methods such as supervised classification (described on this section) or unsupervised classification (discussed on the next section). We present a high-level overview of these methods to explain them in their application to social media during emergencies, the interested reader can consult the material suggested in Section 4.5 for an in-depth exposition.

A supervised classification system is based on a *supervised learning* method, which is a statistical method that creates a general statistical description of a class of items, and/or learns statistical properties that discrimi-

Table 4.1. *(cont.)*

By information source	
Eyewitnesses and/or public	citizen reporters, members of the community (Metaxas and Mustafaraj, 2013); eyewitnesses (Bruns et al., 2012; Diakopoulos et al., 2012; Kumar et al., 2013; Olteanu et al., 2014); local, personally connected (Starbird et al., 2010); local individuals (Starbird et al., 2012; Vieweg et al., 2010); local perspective, on the ground reports (Thomson et al., 2012); direct experience (personal narrative and eyewitness reports) (Shaw et al., 2013); direct observation, direct impact, relayed observation (Truelove et al., 2014); public (St. Denis et al., 2014)
Government	administration/government (Olteanu et al., 2014); police and fire services (Hughes et al., 2014); government (Bruns, 2014); news organization and authorities (Metaxas and Mustafaraj, 2013); public institutions (Thomson et al., 2012); police (Denef et al., 2013); government (Bruns et al., 2012); public service agencies, flood specific agencies (Starbird et al., 2010)
NGOs	nongovernmental organizations (de Choudhury et al., 2012; Olteanu et al., 2014); nonprofit organizations (Thomson et al., 2012); faith-based organizations (Starbird et al., 2010)
News media	news organizations and authorities, blogs (Metaxas and Mustafaraj, 2013), journalists, media, bloggers (de Choudhury et al., 2012); news organizations (Olteanu et al., 2014); professional news reports (Leavitt and Clark, 2014); media (Bruns, 2014); traditional media (print, television, radio), alternative media, freelance journalists (Thomson et al., 2012); blogs, newscrawler bots, local, national, and alternative media (Starbird et al., 2010); media sharing (news media updates, multimedia) (Shaw et al., 2013)
By credibility	
Credible information	credibility (Castillo et al., 2013); credible topics (Canini et al., 2011); content credibility (Gupta and Kumaraguru, 2012); users and content credibility (Gupta et al., 2014); source credibility (Thomson et al., 2012); real images (Gupta et al., 2013)
Rumors	rumor (Hughes et al., 2014; Castillo et al., 2013)
Corrections	rumor mitigation (St. Denis et al., 2014); rumor refutation (Castillo et al., 2013)
By time	
Pre-phase/ preparedness	posted before an actual event occurs, helpful for the preparedness phase of emergency management: pre-disaster, early information (Iyengar et al., 2011; Chowdhury et al., 2013)
Impact-phase/ response	posted during the impact phase of an event, helpful for the response phase of emergency management: during- disaster (Iyengar et al., 2011; Chowdhury et al., 2013)
Post-phase/ recovery	posted after the impact of an event, helpful during the recovery phase: postdisaster information (Chowdhury et al., 2013; Iyengar et al., 2011)
By location	
Ground Zero	information from Ground Zero (victims reports, bystanders) (de Longueville et al., 2009; Ao et al., 2014)
Near-by areas	information originating close to the affected areas (de Longueville et al., 2009)
Outsiders	information coming from other parts of world, sympathizers (Kumar et al., 2013); distant witness (in the sense of Carvin, 2013); location inference (Ikawa et al., 2012); remote crowd (Starbird et al., 2012); nonlocals (Starbird et al., 2010; Thomson et al., 2012)

nate among different classes of items. The process by which these statistical descriptions or models are created automatically is known as statistical *machine learning*.

Four main elements can be identified in a supervised classification system:

(i) the labeled examples from which the statistical model is created, known as training examples;
(ii) the method used to select the dimensions used for representing the elements, known as feature selection;
(iii) the algorithm used for creating the statistical model, known as the learning scheme; and
(iv) the evaluation metrics used to evaluate and report the performance of the system.

Training examples. Supervised learning methods require messages for which the label, that is, the specific information category, is already known. These messages are usually labeled by experts or volunteers. This manual labeling is orders of magnitude slower than what the automatic classification system can achieve, but it may also be considered more precise.

The number of labeled messages required to train a system depends on many factors, including the number of categories into which messages have to be classified, the distribution of messages into those categories, and the variability of messages inside each category. Situations with many categories, with skewed distributions of messages (e.g., categories having very few examples), and with broad categories containing many different kinds of messages, need a relatively larger number of training examples.

In practice, learning a model that can accurately place short messages into one category, requires hundreds or thousands of examples for that category, depending on the desired level of accuracy. More examples yield better results in general, with diminishing results after a certain point (see, e.g., Matykiewicz and Pestian, 2012). Training set sizes reported in the literature for training supervised classifiers on social media or text messages in crises range from hundreds (Yin et al., 2012) to thousands (Imran et al., 2014a) or tens of thousands (Melville et al., 2013) of elements.

As important as the number of training examples, is that they are sampled from the same distribution as the messages that we want to classify. This is a key assumption for many statistical machine learning methods, and indeed it has been observed that the accuracy of models created using training data from one crisis decreases when applied to a different crisis, or when applied to the same crisis but at a different point in time (Imran et al., 2014b). However, given the cost of creating large training sets, there are a series of methods that perform *domain adaptation* or *domain transfer*, which adapt a model created from one dataset to make it useful in another (Dai et al., 2007).

Feature selection. As we described on Section 2.4, messages are converted to a format suitable for algorithms (including machine learning algorithms) by means of feature extraction. Even if messages are brief, and even if aggressive stopword removal, normalization and stemming/lemmatization operations are applied, the feature space in which they can be represented is typically high dimensional. For instance, it could have one dimension for every word plus one dimension for every possible sequence of two words, which for collections of text messages of moderate to large size, quickly yields tens of thousands of dimensions. This may not only increase the amount of computational resources required for the data analysis, but it also increases the chances of overfitting the training data. Conversely, feature selection often yields faster and better classifiers.

Feature selection methods aim at finding a subset of the input features that, for a given purpose, represents the data as well as the entire set of features. Feature selection methods tend to discard features or groups of features that are redundant (e.g., highly correlated with other features) or irrelevant (i.e., not related to the target label).

A simple example of a feature selection criteria is *pointwise mutual information*, in which each single feature is evaluated in isolation in terms of its mutual information with respect to a class.[3] Using this criteria, if a word appears with the same frequency in elements of the training set irrespective of their label, then its mutual information with respect to the label is zero, the word is considered irrelevant and its respective feature discarded.

Feature selection based on analyzing the utility of one feature at a time, and then greedily picking a set made of those features with the largest utility, are not guaranteed to find an optimal subset of features, because of statistical dependencies among features (e.g., two features in isolation may each one be irrelevant, but combined might be relevant). For feature selection methods see, e.g., Guyon and Elisseeff (2003) and references therein.

Learning algorithms. After features have been extracted and selected from the training examples, a machine learning algorithm can be applied. Machine learning algorithms have two modes of operation: learning (or training) and labeling (or testing). The first mode is usually more time-consuming and complex than the second one, which is expected to be fast, given that it will be applied to a large set of data.

There are many supervised classification algorithms that have been used for text classification. They include, among others, naïve Bayes, Support Vector Machines (SVM), logistic regression, decision trees, and random forests; for a survey, see Sebastiani (2002).

The choice of a specific method is to a large extent dependent on the specific problem setting. In most cases, researchers in the literature tend to test two or three different algorithms, and then decide which one is better for their problem setting given their evaluation metrics. The choice of an algorithm may also depend on practicalities such as the availability of an efficient and robust implementation in a given programming language or a given platform. For instance, *ESA* (Yin et al., 2012; Cameron et al., 2012) uses naïve Bayes and SVM; *EMERSE* (Caragea et al., 2011) and Neubig et al. (2011) use SVM; *AIDR* (Imran et al., 2014a) uses random forests; *Tweedr* (Ashktorab

[3] The pointwise mutual information between a term t and a category c is $\mathrm{pmi}(t, c) = \log(p(t, c)/p(t)p(c))$ where $p(t, c)$ is the probability of a message belonging to the category and containing the term, $p(t)$ is the probability of a message containing the term, and $p(c)$ is the probability of a message belonging to a category.

et al., 2014) uses logistic regression. While in most cases algorithms are used to predict a single label for each element, adaptations of these algorithms for the multilabel case are sometimes employed (e.g., Caragea et al., 2011).

Supervised learning methods typically produce parametrized models, in which the only difference between two models is a vector of parameters, coefficients, or weights. In general, learning algorithms are often used as "black boxes" in which those parameters do not need to be inspected. However, there are cases in which we seek to understand how specific aspects of the input are related to the output label. Verma et al. (2011), for instance, observe that the messages that contribute the most to situational awareness are also those that are expressed using objective (as opposed to subjective) language. Following this observation, they create a *stacked classifier* in which at one level certain characteristics of the message are modeled (e.g., by having a classifier that classifies messages as objective or subjective) and at the next level these characteristics are combined with other characteristics (e.g., writing styles such as formal or informal), and with features from the message itself. Verma et al. (2011) find that this approach performs better than directly using the input features on their dataset.

Along the same lines, in some cases there is background knowledge that can guide the creation of the model. For instance, Melville et al. (2013) use the pooling multinomials method (Melville et al., 2009). Pooling multinomials takes as input labels on items *and on features*. Following the example from the beginning of this section, if a priori we know that the word "shelter" is a good indicator of the category "emergency shelter," then instead of adding a hard rule that the word "shelter" automatically implies this class, we can pass this information to the learning algorithm to use it as part of the creation of the statistical models for the different classes.

Evaluation metrics. Comparing different classification systems is not a trivial task. With some exceptions, there are few reference collections in which to perform comparative evaluations.[4] As other information processing operations, classification is usually measured in terms of efficiency and effectiveness. Efficiency in this case is basically a matter of the speed of the system. Effectiveness can be measured in a number of ways, and depends on many factors, including the training data, learning scheme, target categories, and in some cases the language for which a classifier is built Zielinski et al. (2012).

A key concept to understand the effectiveness of an automatic classification system is its *confusion matrix*, which is a table representing correct and incorrect classifications. For instance, let's assume a classifier has been created to distinguish between messages containing information about "people," "infrastructure," or "other," and that we have external information that allows us to validate messages that actually belong to each of these categories (e.g., expert assessments over a subset of the data). In this case, the confusion matrix would be a 3×3 numerical matrix with the following structure:

	Actually about people	Actually about infrastructure	Actually about other
Classified as "people"	Pp	Ip	Op
Classified as "infrastructure"	Pi	Ii	Oi
Classified as "other"	Po	Io	Oo

Accuracy is probably the simpler metric for classification effectiveness. It corresponds to the probability that an item is classified correctly: $(Pp + Ii + Oo)/(Pp + Pi + Po + Ip + Ii + Io + Op + Oi + Oo)$. It is the sum of the values in the diagonal of the confusion matrix, divided by the sum of all cells. Classification accuracies reported in the literature of social media during crises range from 0.6 to 0.9 (Imran et al., 2015). Accuracy can be misleading

[4] CrisisLex: collections and lexicons for analyzing crisis-related social media. http://crisislex .org/.

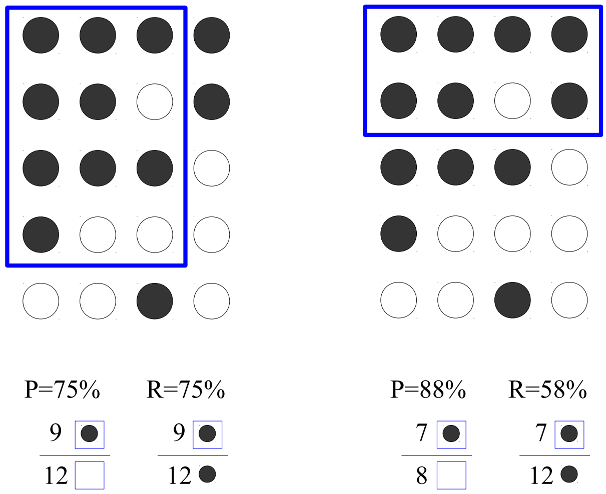


Figure 4.1. Visual depiction of precision and recall, comparing two hypothetical systems. Messages are represented by circles, and the messages classified as being relevant for a given class are depicted by large rectangles. Messages that are actually relevant are filled/black; irrelevant messages are empty/white. The system on the right has more precision (it has only one false positive), but it has lower recall (it has five false negatives).

when dealing with imbalanced classes. For instance, suppose there are 100 messages in total, out of which one is about people, one is about infrastructure, and 98 are about other aspects. In this case, a classifier that always says "other" has $Po = 1, Io = 1, Oo = 98$ and all other values 0, which yields an accuracy of 98%, even if the classifier is blindly outputting always the same category.

Precision and recall are measures that can deal better with imbalanced cases. *Precision* (P) is a measure of specificity. It corresponds to the probability that an item that we have classified as belonging to a class, actually belongs to that class. Precision is measured on a per-class basis. The precision for the class "people" in the previous example would be: $Pp/(Pp + Ip + Op)$, that is, out of everything that was classified as "people," the fraction that was actually about people. *Recall* (R) is a measure of sensitivity. Recall is also measured on a per-class basis. It corresponds to the probability that an item that actually belongs to a class is classified by us as belonging to that class. The recall for the class "people" in the example given earlier, would be $Pp/(Pp + Pi + Po)$, that is, out of every element that was about people, the fraction that we actually classified as "people." A visual depiction of precision and recall is shown in Figure 4.1.

A popular metric combining both is their harmonic mean, known as the F_1 measure or simply F measure:

$$F = F_1 = 2\frac{PR}{P + R} .$$

A more general metric for combining precision and recall is the F_β measure, defined as:

$$F_\beta = (1 + \beta^2)\frac{PR}{\beta^2 P + R} .$$

By varying the parameter β we can achieve different trade-offs. For instance, F_2 favors recall at the expense of precision, and $F_{0.5}$ favors precision at the expense of recall.

Some supervised machine learning methods generate an output that is not a category label, but instead a score which is related to the probability that an element belongs to a class (or a set of classes). In this case, the score is usually thresholded at a particular value (e.g., zero) so that items with scores above the threshold are labeled as belonging to a class, and the items with scores below the threshold are labeled as not belonging to it. In these cases, there is more than once choice of threshold, which yields different trade-offs of precision and recall. In order to compare two systems irrespectively of the chosen threshold, a standard approach is to compute the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve maps the performance of a classifier for every choice of threshold to a point in the plane, in which the false positive rate obtained with that threshold goes in the X axis, and the true positive rate in the Y axis. When the area under the curve (AUC) in the interval

[0, 1] is 0.5, the classifier is not better than a random classifier, and when it is 1.0, the classifier is perfect. For an overview on the usage of ROC curves to compare classification systems, see Fawcett (2004).

4.3 Unsupervised Classification / Clustering

Clustering means grouping similar items together. Clustering is an *unsupervised machine learning method*, which is a large class of exploratory methods that search for patterns or structure in unlabeled data.

The input to a clustering algorithm is a set of items, in our case, vectors representing messages, and a way of measuring how similar two messages are, that is, a similarity function. A frequently used similarity function for text is the *cosine similarity* (see, e.g., Baeza-Yates and Ribeiro-Neto, 2011, ch. 6). In most cases, the input should also include the desired number of classes, although there are clustering methods that can apply some criterion to determine a "good" number of classes.

The output of a clustering algorithm is a mapping from items to classes, in which items in the same class are expected to be similar to each other, and items in different classes are expected to be different from each other. Depending on the specific algorithm used, the classes do not need to be disjoint, that is, a message can belong to more than one class. In this case, we speak of *soft clustering*, in contrast with *hard clustering*, where each message must belong to one and only one class.

Example hard clustering method: k-means. A popular clustering algorithm for documents is *k-means*, which is a centroid-based method. This algorithm operates by iterating between two steps. For the initialization, items can be assigned randomly to classes. In the first step, the algorithm computes the *centroid* of each class of items, that is, the item that minimizes the average distance to all of the items on that class. In the second step, documents are reassigned to the class whose centroid is most similar to them. The first and second step are repeated in sequence a number of times, until a certain stop criterion is reached. For instance, it can be done for a fixed number of iterations, or until the average distances to centroids stops decreasing, or do not decrease more than a certain amount per iteration. There are many other clustering algorithms in addition to k-means; for an overview, see Zaki and Meira (2014, part III).

Example soft clustering method: LDA. Latent Dirichlet Allocation (LDA), introduced by Blei et al. (2003), is often used to create a soft clustering of documents into a predefined number of topics. It assumes that every document reflects a combination of topics, in which the number of relevant topics for a document are a relatively small fraction of all the possible topics. It also assumes that every topic can be characterized by a small set of characteristic words that are highly probable for that topic, and that most words have the same probability across all topics.

LDA is a probabilistic model that considers documents as a result of a probabilistic process. Each document is defined by a probability distribution over topics, and each topic is defined as a probability distribution over words. To create each word in a document, we just need to sample from the topics according to the topic-distribution of that document, and then sample from the words of that topic according to that topic word-distribution.

The "L" in LDA stands for latent, which reflects the fact that while we can observe directly the words in each document, we do not know the topic-distributions or the word-distributions. These distribution need to be estimated using a probabilistic estimation method, such as Gibbs sampling (Wei and Croft, 2006). The "D" stands for the Dirichlet distribution, which is used to incorporate an assumption of sparsity, that is, that each document has few relevant topics and that each topic has few characteristic words. LDA is closely related to a method in Information Retrieval known as Probabilistic Latent Semantic Indexing (PLSI), described by Hofmann (1999).

The output of LDA is a probability distribution over topics for each of the documents. This distribution can be thresholded to read it as a soft clustering assignment, for instance, every document having a probability larger than a certain threshold of belonging to a topic, is assigned to that topic. Interestingly, words that have a high probability in the word distribution of a topic can be used as human-readable "summmaries" for the topic.

Clustering granularity. In the context of social media messages during crises, there are two prototypical scenarios in which clustering can be used. In some cases, we want to perform clustering to group together messages that refer to the same aspect of the crisis. In this case, typically a few large clusters are created, which we call the coarse-granularity setting. In other cases, we want to perform clustering to group together messages that convey basically the same information. In this case, typically many small clusters are created, which we call the fine-granularity setting.

A *coarse-granularity setting* (few large clusters) is used by Kireyev et al. (2009), which applies LDA to data from a 2009 Earthquake in Indonesia. The output of LDA includes broad topics that cover different aspects of the crisis such as one topic in which high-probability words are {*tsunami*, *disaster*, *relief*, *earthquake*}, and other topics represented by words {*dead*, *bodies*, *missing*, *victims*} and {*aid*, *help*, *money*, *relief*}. Nelson and Pottenger (2013) describe the usage of LDA for a similar application involving SMS messages in a coarse-granularity setting. Karandikar (2010) shows that LDA can be applied to compare data across different disasters (eight in their case), recovering important keywords which are characteristic of each disaster. Coarse-granularity clustering is also used for event detection, for example, by Berlingerio et al. (2013) and others, as described in Section 6.4.

A *fine-granularity setting* (many small clusters) can be used to help reduce the number of social media messages that need to be processed/examined by humans, for instance, by displaying multiple equivalent messages as a single item instead of multiple ones. This is the approach used by *CrisisTracker* (Rogstadius et al., 2013), which is a crowdsourced social media curation system for disaster awareness. The system, which collects data from Twitter based on predefined filters (i.e., keywords, bounding box), groups these messages into many *stories*, which are small clusters of tweets. These stories are then curated/classified by humans, which is much more efficient than repeatedly classifying messages that are almost equivalent to each other. The specific clustering method employed in this case is Locality-Sensitive Hashing (LSH), an efficient technique that uses hash functions to detect near-duplicates in data (Charikar, 2002).

4.4 Research Problems

Adapting/transferring classification models to new situations. A recurrent theme in the sociology of disaster literature is that, despite their superficial differences, disasters tend to have many elements in common with each other. Being able to reuse human-labeled and classification models can be very useful to produce results early on when a new disaster strikes. Indeed, Li et al. (2015) describe a domain adaptation approach that can be applied to Twitter data for disaster response.

Performing interactive taxonomy design. Designing content categories is an art that is difficult to master. Furthermore, sudden-onset disasters and emergency situations do not leave enough time to spend on creating new categorizations. The main source of uncertainty is that, while we may guess a priori which topics will appear in social media, the relative prominence of different topics is hard or impossible to estimate. Adaptive, interactive approaches in which experts interact with algorithms, data, and annotators to rapidly converge into an appropriate typology, could be very useful to accelerate the construction of an appropriate typology for a crisis.

Ranking. After categorizing messages, some categories may be very large; in this case one can attempt to summarize the category or to pick a few representatives of it. One form of choosing such representatives is by ranking, for example, by sorting elements in decreasing order of importance. Ranking is a very difficult problem in Information Retrieval; on the crisis domain, there are relatively few works dealing with ranking (Li et al., 2012).

4.5 Further Reading

For a high-level data mining perspective, Zaki and Meira (2014, parts III and IV) present algorithms for clustering and classification.

Ingersoll et al. (2013, ch. 6 and 7) focuses on text clustering and classification from a practical/practitioner perspective. Baeza-Yates and Ribeiro-Neto (2011, ch. 8) describes various algorithms on text classification, including aspects of feature selection and how to evaluate text classification methods. Joachims (2002) describes the general setting of text classification in its initial chapters, then presents in detail how to use Support Vector Machines (SVMs) for classifying text.

Many of the methods described on this section can be applied to other types of data beyond brief text messages, including blog and news content (Leetaru and Schrodt, 2014).

Bibliography

Acar, Adam, and Muraki, Yuya. 2011. Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, **7**(3), 392–402.

Ao, Ji, Zhang, Peng, and Cao, Yanan. 2014. Estimating the locations of emergency events from Twitter streams. *Procedia Computer Science*, **31**, 731–739.

Ashktorab, Zahra, Brown, Christopher, Nandi, Manojit, and Culotta, Aron. 2014. Tweedr: Mining Twitter to inform disaster response. In: *Proceedings of 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. University Park, Pennsylvania, USA: ISCRAM.

Baeza-Yates, Ricardo, and Ribeiro-Neto, Berthier. 2011. *Modern information retrieval: The concepts and technology behind search*. 2nd edn. Addison-Wesley Professional.

Barrenechea, Mario, Anderson, Kenneth M., Palen, Leysia, and White, Joanne. 2015. Engineering crowdwork for disaster events: The human-centered development of a lost-and-found tasking environment. Pages 182–191 of: *Proceedings of 48th Annual Hawaii International Conference on System Sciences (HICSS)*. Kauai, Hawaii, USA: IEEE, for IEEE.

Berlingerio, Michele, Francesco, Calabrese, Giusy, Di Lorenzo, Xiaowen, Dong, Yiannis, Gkoufas, and Mavroeidis, Dimitrios. 2013. SaferCity: A system for detecting and analyzing incidents from social media. Pages 1077–1080 of: *Proceedings of 13th International Conference on Data Mining Workshops ICDMW*. Dallas, Texas, USA: IEEE.

Blei, David M., Ng, Andrew Y., and Jordan, Michael I. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, **3**, 993–1022.

Bruns, Axel. 2014. Crisis communication. *The Media and Communications in Australia*, 351–355.

Bruns, Axel, Burgess, Jean E, Crawford, Kate, and Shaw, Frances. 2012. *#qldfloods and @qpsmedia: Crisis communication on Twitter in the 2011 south east Queensland floods*. Tech. rept. ARC Centre, Queensland University of Technology.

Cameron, Mark A., Power, Robert, Robinson, Bella, and Yin, Jie. 2012. Emergency situation awareness from Twitter for crisis management. Pages 695–698 of: *Social Web for Disaster Management (SWDM), Companion: Proceedings of 21st International Conference on World Wide Web Conference (WWW)*. Lyon, France: ACM.

Canini, Kevin Robert, Suh, Bongwon, and Pirolli, Peter L. 2011. Finding credible information sources in social networks based on content and social structure. Pages 1–8 of: *Proceedings of IEEE 3rd International Conference on Privacy, Security, Risk and Trust PASSAT*. Boston, Massachusetts, USA: IEEE.

Caragea, Cornelia, McNeese, Nathan, Jaiswal, Anuj, Traylor, Greg, Kim, H., Mitra, Prasenjit, Wu, Dinghao, Tapia, A., Giles, Lee, Jansen, Bernard J., et al. 2011. Classifying text messages for the Haiti earthquake. In: *Proceedings of 8th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Lisbon, Portugal: ISCRAM.

Carvin, Andy. 2013. *Distant witness*. City University of New York CUNY Journalism Press.

Castillo, Carlos, Mendoza, Marcelo, and Poblete, Barbara. 2013. Predicting information credibility in time-sensitive social media. *Internet Research*, **23**(5), 560–588.

Charikar, Moses S. 2002. Similarity estimation techniques from rounding algorithms. Pages 380–388 of: *Proceedings of 34th Annual ACM Symposium on Theory of Computing*. Montreal, Quebec, Canada: ACM.

De Choudhury, Munmun, Diakopoulos, Nicholas, and Naaman, Mor. 2012. Unfolding the event landscape on Twitter: Classification and exploration of user categories. Pages 241–244 of: *Proceedings of ACM Conference on Computer Supported Cooperative Work (CSCW)*. Seattle, Washington, USA: ACM.

Chowdhury, Soudip Roy, Imran, Muhammad, Asghar, Muhammad Rizwan, Amer-Yahia, Sihem, and Castillo, Carlos. 2013. Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In: *Proceedings of 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Baden Baden, Germany: ISCRAM.

Dai, Wenyuan, Xue, Gui-Rong, Yang, Qiang, and Yu, Yong. 2007. Transferring naive bayes classifiers for text classification. Pages 540–545 of: *Proceedings of 22nd AAAI Conference on Artificial Intelligence*, vol. 22. Vancouver, Canada: AAAI Press.

Denef, Sebastian, Bayerl, Petra S., and Kaptein, Nico A. 2013. Social media and the police: Tweeting practices of British police forces during the August 2011 riots. Pages 3471–3480 of: *Proceedings of ACM Conference on Human Factors in Computing Systems (SIGCHI)*. Paris, France: ACM.

Diakopoulos, Nicholas, De Choudhury, Munmun, and Naaman, Mor. 2012. Finding and assessing social media information sources in the context of journalism. Pages 2451–2460 of: *Proceedings of Conference on Human Factors in Computing Systems (SIGCHI)*. Austin, Texas, USA: ACM.

Fawcett, Tom. 2004. Roc graphs: Notes and practical considerations for researchers. *Machine Learning*, **31**, 1–38.

Fischer, Henry W. 1998. *Response to disaster: Fact versus fiction and its perpetuation: the sociology of disaster*. University Press of America.

Gupta, Aditi, and Kumaraguru, Ponnurangam. 2012. Credibility ranking of tweets during high impact events. In: *Proceedings of 1st Workshop on Privacy and Security in Online Social Media PSOSM, at WWW 2012*. Lyon, France: ACM.

Gupta, Aditi, Lamba, Hemank, Kumaraguru, Ponnurangam, and Joshi, Anupam. 2013. Faking Sandy: Characterizing and identifying fake images on Twitter during Hurricane Sandy. Pages 729–736 of: *Privacy and Security in Online Social Media (PSOSM), Companion: Proceedings of 22nd International Conference on World Wide Web Conference (WWW)*. Rio de Janeiro, Brazil: ACM, for IW3C2.

Gupta, Aditi, Kumaraguru, Ponnurangam, Castillo, Carlos, and Meier, Patrick. 2014. Tweetcred: Real-time credibility assessment of content on Twitter. Pages 228–243 of: *Proceedings of 1st Workshop on Privacy and Security in Online Social Media SocInfo*. Barcelona, Spain: Springer.

Guyon, Isabelle, and Elisseeff, André. 2003. Special issue on variable and feature selection. *Journal of Machine Learning Research*, **3**(Mar.), 1157–1461.

Hofmann, Thomas. 1999 (Aug.). Probabilistic latent semantic indexing. Pages 50–57 of: *Proceedings of 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*. ACM, Berkeley, California, USA.

Hughes, Amanda L., St. Denis, Lise Ann, Palen, Leysia, and Anderson, Kenneth. 2014. Online public communications by police and fire services during the 2012 Hurricane Sandy. Pages 1505–1514 of: *Proceedings of ACM Conference on Human Factors in Computing Systems (SIGCHI)*. Toronto, Canada: ACM.

Ikawa, Yohei, Enoki, Miki, and Tatsubori, Michiaki. 2012. Location inference using microblog messages. Pages 687–690 of: *Proceedings of 21st International Conference on World Wide Web Conference (WWW)*. Lyon, France: ACM.

Imran, Muhammad, Elbassuoni, Shady Mamoon, Castillo, Carlos, Diaz, Fernando, and Meier, Patrick. 2013. Extracting information nuggets from disaster-related messages in social media. In: *Proceedings of 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Baden Baden, Germany: ISCRAM.

Imran, Muhammad, Castillo, Carlos, Lucas, Ji, Meier, Patrick, and Vieweg, Sarah. 2014a. AIDR: Artificial Intelligence for Disaster Response. Pages 159–162 of: *Social Web for Disaster Management (SWDM), Companion: Proceedings of 23rd International Conference on World Wide Web (WWW)*. Seoul, Korea: ACM, for IW3C2.

Imran, Muhammad, Castillo, Carlos, Lucas, Ji, Patrick, M, and Rogstadius, Jakob. 2014b. Coordinating human and machine intelligence to classify microblog communications in crises. In: *Proceedings of 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. University Park, Pennsylvania, USA: ISCRAM.

Imran, Muhammad, Castillo, Carlos, Diaz, Fernando, and Vieweg, Sarah. 2015. Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, **47**(4), 67:1–67:38.

Ingersoll, Grant S., Morton, Thomas S., and Farris, Andrew L. 2013. *Taming text: How to find, organize, and manipulate it*. 1st edn. Manning Publications.

Iyengar, Akshaya, Finin, Tim, and Joshi, Anupam. 2011. Content-based prediction of temporal boundaries for events in Twitter. Pages 186–191 of: *Proceedings of IEEE 3rd International Conference on Privacy, Security, Risk and Trust PASSAT*. Boston, Massachusetts, USA: IEEE.

Joachims, Thorsten. 2002. *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Karandikar, Anand. 2010. *Clustering short status messages: A topic model based approach*. M.Phil. thesis, University of Maryland.

Kireyev, Kirill, Palen, Leysia, and Anderson, K. 2009. Applications of topics models to analysis of disaster-related Twitter data. In: *NIPS Workshop on Applications for Topic Models: Text and beyond*, vol. 1.

Kumar, Shamanth, Morstatter, Fred, Zafarani, Reza, and Liu, Huan. 2013. Whom should I follow? Identifying relevant users during crises. Pages 139–147 of: *Proceedings of 24th ACM Conference on Hypertext and Social Media Hypertext*. Paris, France: ACM.

Leavitt, Alex, and Clark, Joshua A. 2014. Upvoting hurricane Sandy: Event-based news production processes on a social news site. Pages 1495–1504 of: *Proceedings of Conference on Human Factors in Computing Systems (SIGCHI)*. Toronto, Canada: ACM.

Leetaru, Kalev, and Schrodt, Philip A. 2014. *GDELT: Global Database of Events, Language, and Tone*. Online Website.

Li, Hongmin, Guevara, Nicolais, Herndon, Nic, Caragea, Doina, Neppalli, Kishore, Caragea, Cornelia, Squicciarini, Anna, and Tapia, Andrea H. 2015. Twitter mining for disaster response: A domain adaptation approach. In: *Proceedings of 12th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Kristiansand, Norway: ISCRAM.

Li, Rui, Lei, Kin Hou, Khadiwala, Ravi, and Chang, KC-C. 2012 (Apr.). Tedas: A Twitter-based event detection and analysis system. Pages 1273–1276 of: *Proceedings of 28th International Conference on Data Engineering ICDE*. IEEE, Washington DC, USA.

De Longueville, Bertrand, Smith, Robin S., and Luraschi, Gianluca. 2009. OMG, from here, I can see the flames! A use case of mining location based social networks to acquire spatio-temporal data on forest fires. Pages 73–80 of: *Proceedings of International Workshop on Location Based Social Networks LBSN*. Washington DC, USA: ACM, for ACM.

Matykiewicz, Pawel, and Pestian, John. 2012. Effect of small sample size on text categorization with support vector machines. Pages 193–201 of: *Proceedings of 2012 Workshop on Biomedical Natural Language Processing*. Montreal, Canada: ACL, for Association for Computational Linguistics.

Melville, Prem, Gryc, Wojciech, and Lawrence, Richard D. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. Pages 1275–1284 of: *Proceedings of 15th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Paris, France: ACM, for ACM.

Melville, Prem, Chenthamarakshan, Vijil, Lawrence, Richard D., Powell, James, Mugisha, Moses, Sapra, Sharad, Anandan, Rajesh, and Assefa, Solomon. 2013. Amplifying the voice of youth in Africa via text analytics. Pages 1204–1212 of: *Proceedings of 19th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. Chicago, Illinois, USA: ACM.

Metaxas, Panagiotis, and Mustafaraj, Eni. 2013. The rise and the fall of a citizen reporter. Pages 248–257 of: *Proceedings of 5th Annual ACM Web Science Conference WebSci*. Paris, France: ACM.

Mileti, Dennis. 1999. *Disasters by design: A reassessment of natural hazards in the United States*. Natural Hazards and Disasters: Reducing Loss and Building Sustainability in a Hazardous World. Joseph Henry Press.

Nelson, Christie, and Pottenger, William M. 2013. Optimization of emergency response using higher order learning and clustering of 911 text messages. Pages 486–491 of: *Technologies for Homeland Security (HST), 2013 IEEE International Conference on*. IEEE.

Neubig, Graham, Matsubayashi, Yuichiroh, Hagiwara, Masato, and Murakami, Koji. 2011. Safety information mining-what can nlp do in a disaster. Pages 965–973 of: *Proceedings of 5th International Joint Conference on Natural Language Processing (IJCNLP)*, vol. 11. Chiang Mai, Thailand: ACL.

Olteanu, Alexandra, Castillo, Carlos, Diaz, Fernando, and Vieweg, Sarah. 2014 (June). CrisisLex: A lexicon for collecting and filtering microblogged communications in crises. In: *Proceedings of 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Olteanu, Alexandra, Vieweg, Sarah, and Castillo, Carlos. 2015. What to expect when the unexpected happens: Social media communications across crises. Pages 994–1009 of: *Proceedings of Conference on Computer-Supported Cooperative Work (CSCW)*. Vancouver, Canada: ACM.

Parsons, Sophie, Atkinson, Peter M., Simperl, Elena, and Weal, Mark. 2015. Thematically analysing social network content during disasters through the lens of the disaster management lifecycle. In: *Proceedings of Social Web for Disaster Management (SWDM) at WWW 2015*. Florence, Italy: ACM.

Petak, William J. 1985. Emergency management: A challenge for public administration. *Public Administration Review*, 3–7.

Purohit, Hemant, Castillo, Carlos, Diaz, Fernando, Sheth, Amit, and Meier, Patrick. 2014a. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday Online Journal*, **19**(1).

Purohit, Hemant, Hampton, Andrew, Bhatt, Shreyansh, Shalin, Valerie L., Sheth, Amit P., and Flach, John M. 2014b. Identifying seekers and suppliers in social media communities to support crisis coordination. *Computer Supported Cooperative Work (CSCW)*, **23**(4-6), 513–545.

Qu, Yan, Huang, Chen, Zhang, Pengyi, and Zhang, Jun. 2011. Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. Pages 25–34 of: *Proceedings of Conference on Computer-Supported Cooperative Work (CSCW)*. Hangzhou, China: ACM.

Rogstadius, Jakob, Vukovic, Maja, Teixeira, Claudio, Kostakos, Vassilis, Karapanos, Evangelos, and Laredo, Jim Alain. 2013. Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, **57**(5), 4–1.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, **34**(1), 1–47.

Shaw, Frances, Burgess, Jean, Crawford, Kate, and Bruns, Axel. 2013. Sharing news, making sense, saying thanks: Patterns of talk on Twitter during the Queensland floods. *Australian Journal of Communication*.

Smith, Luke, Liang, Qiuhua, James, Phil, and Lin, Wen. 2015. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management*, Jan.

Sreenivasan, Nirupama Dharmavaram, Lee, Chei Sian, and Goh, Dion Hoe-Lian. 2011. Tweet me home: Exploring information use on Twitter in crisis situations. Pages 120–129 of: *Online Communities and Social Computing*, vol. 6778. Springer-Verlag Berlin Heidelberg. ISBN 978-3-642-21796-8.

St. Denis, Lise Ann, Palen, Leysia, and Anderson, Kenneth M. 2014. Mastering social media: An analysis of Jefferson county's communications during the 2013 Colorado Floods. In: *Proceedings of 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. University Park, Pennsylvania, USA: ISCRAM.

Starbird, Kate, Palen, Leysia, Hughes, Amanda L, and Vieweg, Sarah. 2010. Chatter on the red: What hazards threat reveals about the social life of microblogged information. Pages 241–250 of: *Proceedings of ACM Conference on Computer-Supported Cooperative Work (CSCW)*. Savannah,Georgia, USA: ACM, for ACM.

Starbird, Kate, Muzny, Grace, and Palen, Leysia. 2012. Learning from the crowd: Collaborative filtering techniques for identifying on-the-ground twitterers during mass disruptions. In: *Proceedings of 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Vancouver, Canada: ISCRAM.

Taylor, Mel, Wells, Garrett, Howell, Gwyneth, and Raphael, Beverley. 2012. The role of social media as psychological first aid as a support to community resilience building. *Australian Journal of Emergency Management, The*, **27**(1), 20–26.

Thomson, Robert, Ito, Naoya, Suda, Hinako, Lin, Fangyu, Liu, Yafei, Hayasaka, Ryo, Isochi, Ryuzo, and Wang, Zian. 2012. Trusting tweets: The Fukushima disaster and information source credibility on Twitter. Page 10 of: *Proceedings of 9th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*. Vancouver, Canada: ISCRAM.

Truelove, Marie, Vasardani, Maria, and Winter, Stephan. 2014. Towards credibility of micro-blogs: characterising witness accounts. *GeoJournal*, 1–21.

Verma, Sudha, Vieweg, Sarah, Corvey, William J, Palen, Leysia, Martin, James H, Palmer, Martha, Schram, Aaron, and Anderson, Kenneth Mark. 2011 (July). Natural language processing to the rescue? extracting "situational awareness" tweets during mass emergency. In: *Proceedings of 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Vieweg, Sarah. 2012. *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. Ph.D. thesis, University of Colorado, Boulder.

Vieweg, Sarah, Hughes, Amanda L, Starbird, Kate, and Palen, Leysia. 2010. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. Pages 1079–1088 of: *Proceedings of Conference on Human Factors in Computing Systems (SIGCHI)*. Atlanta, Georgia, USA: ACM, for ACM.

Wei, Xing, and Croft, W Bruce. 2006. LDA-based document models for ad-hoc retrieval. Pages 178–185 of: *Proceedings of 29th International Conference on Research and Development in Information Retrieval (SIGIR)*. Seattle, Washington, USA: ACM, for ACM.

White, Joanne I., and Palen, Leysia. 2015. Expertise in the wired wild west. Pages 662–675 of: *Proceedings of 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. Vancouver, Canada: ACM, for ACM.

White, Joanne I., Palen, Leysia, and Anderson, Kenneth M. 2014. Digital mobilization in disaster response: The work and self-organization of on-line pet advocates in response to Hurricane Sandy. Pages 866–876 of: *Proceedings of 17th ACM Conference on Computer Supported Cooperative Work and Social Computing*. Baltimore, Maryland, USA: ACM, for ACM.

Wukich, Clayton, and Mergel, Ines A. 28/08/2014. Closing the citizen-government communication gap: Content, audience, and network analysis of government tweets. *Social Science Research Network Working Paper Series*, Aug., 35 pages.

Yin, Jie, Lampert, Andrew, Cameron, Mark, Robinson, Bella, and Power, Robert. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*, **27**(6), 52–59.

Zaki, Mohammed J., and Meira, Wagner. 2014. *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge University Press.

Zielinski, Andrea, Bügel, Ulrich, Middleton, L, Middleton, SE, Tokarchuk, L, Watson, K, and Chaves, F. 2012. Multilingual analysis of Twitter news in support of mass emergency events. Page 8085 of: *EGU General Assembly Conference Abstracts*, vol. 14.

If you enjoyed this free preview of Big Crisis Data,
get the book at http://bigcrisisdata.org/