CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises

Alexandra Olteanu[*]
Ecole Polytechnique Federale de Lausanne
alexandra.olteanu@epfl.ch

Carlos Castillo
Qatar Computing Research Institute
chato@acm.org

Fernando Diaz
Microsoft Research
fdiaz@microsoft.com

Sarah Vieweg
Qatar Computing Research Institute
svieweg@qf.org.qa

Abstract

Locating timely, useful information during crises and mass emergencies is critical for those forced to make potentially life-altering decisions. As the use of Twitter to broadcast useful information during such situations becomes more widespread, the problem of finding it becomes more difficult. We describe an approach toward improving the recall in the sampling of Twitter communications that can lead to greater situational awareness during crisis situations. First, we create a lexicon of crisis-related terms that frequently appear in relevant messages posted during different types of crisis situations. Next, we demonstrate how we use the lexicon to automatically identify new terms that describe a given crisis. Finally, we explain how to efficiently query Twitter to extract crisis-related messages during emergency events. In our experiments, using a *crisis lexicon* leads to substantial improvements in terms of recall when added to a set of crisis-specific keywords manually chosen by experts; it also helps to preserve the original distribution of message types.

1 Introduction

The popular microblogging platform Twitter is a frequent destination for affected populations during mass emergencies. Twitter is a place to exchange information, ask questions, offer advice, and otherwise stay informed about the event. Those affected require timely, relevant information; recent research shows that information broadcast on Twitter can lead to enhanced situational awareness, and help those faced with an emergency to gain valuable information (Vieweg 2012).

The velocity and volume of messages (*tweets*) in Twitter during mass emergencies makes it difficult to locate situational awareness information, such as road closure locations, or where people need water. Users often employ conventional markers known as *hashtags* to bring attention to specific tweets. The idea is that those looking for emergency information will search for specific hashtags, and tweets that contain the hashtag will be located. In crisis, hashtags are often adopted by an information propagation process (Starbird and Palen 2011), but in some cases, they are suggested by emergency response agencies or other authorities. Alas, even with several dozen such hashtags, only a fraction of the information broadcast on Twitter during mass emergencies is covered (Bruns et al. 2012). Therefore, automatic methods are necessary to help humans cull through the masses of Twitter data to find useful information.

Here, we tackle the problem of how to locate tweets that contain crisis-relevant information during mass emergency situations: our goal is to improve query methods, and return more relevant results than is possible using conventional manually-edited keywords or location-based searches.

Problem definition. Given a crisis situation that occurs within a geographical boundary, automatically determine a query of up to K terms that can be used to sample a large set of crisis-related messages from Twitter.

Our approach. Create a *crisis lexicon* consisting of crisis-related terms that tend to frequently appear across various crisis situations. This lexicon has two main applications:

1. Increase the recall in the sampling of crisis-related messages (particularly at the start of the event), without incurring a significant loss in terms of precision.
2. Automatically identify the terms used to describe a crisis by employing pseudo-relevance feedback mechanisms.

Our approach is presented with respect to crises, but it can be applied to any domain. We describe a systematic method to build the lexicon using existing data samples and crowdsourced labeling; the method is general and can be applied to other tasks (e.g. to build a sports-related or a health-related lexicon). The lexicon, along with the data and the code we used to build it are available at http://crisislex.org/.

2 Related Work

Mining social media in crises. During crises, numerous disaster-related messages are posted to microblogging sites, which has led to research on understanding social media use in disasters (Starbird and Palen 2010; Qu et al. 2011), and extracting useful information (Imran et al. 2013).

The first challenge in using microblog data is to retrieve comprehensive sets of disaster-related tweets (Bruns and Liang 2012). This is due to Twitter's public API limitations (described in §3.1) that make this type of data collection difficult. To the best of our knowledge, data collection during crises usually falls in two categories: keyword-based and location-based, with the former being more common. In a keyword-based collection, a handful of terms and/or hashtags

[*]Work done at the Qatar Computing Research Institute.

are used to retrieve tweets containing those terms (Hughes and Palen 2009) ignoring other posts (Bruns and Liang 2012). While the resulting samples might have little noise (Vieweg et al. 2010), they are typically constructed around visible topical hashtags and might omit a significant number of disaster-related tweets (Bruns et al. 2012). Furthermore, keywords are only as responsive as the humans curating them and this method may lose relevant tweets due to latency. Location-based sampling, on the other hand, is limited to tweets that are either geo-tagged or mention the places affected by the disaster; both of these conditions occur in a small portion of tweets.

Once collected, it is necessary to process the data in a meaningful way. Imran et al. (2013) automatically identify tweets contributing to situational awareness and classify them according to several types of information. Yin et al. (2012) designed a system for leveraging microblog data during disasters; their data capture module is close in scope with our work, yet it makes no distinction between disasters and other events. In turn, our lexicon could enhance their burst detection mechanisms to better identify disasters.

Query generation and expansion. Our problem resembles *deep-web crawling*, the process by which web crawlers access public data (belonging to large online retailers, libraries, etc.) on the web that is not accessible by following links, but only by filling in search forms. To this end, it performs *query generation*: identify a set of keywords that are entered in search forms to return such data (Wu et al. 2006; Ntoulas, Pzerfos, and Cho 2005).

The goal of exhaustively retrieving documents hidden behind web interfaces has been approached as a *minimum weighted dominating set* and *set-covering* graph problem (Ntoulas, Pzerfos, and Cho 2005; Wu et al. 2006). We reuse the idea of representing document or term co-occurrences as a graph, but we formalize our problem as finding the *maximum weighted independent set* as we look for *discriminative* queries that maximize only the volume of retrieved documents *relevant* to given topics (§4.1). In web search, reformulating the initial query such that it returns documents from the domain of interest is known as *vertical selection & aggregation* (Arguello, Diaz, and Paiement 2010). Arguello et al. reuse past knowledge to predict models for new domains by focusing on *portability* and *adaptability*. We use their idea of supervision and use knowledge on past crises to generate queries for future ones.

The query generation step can be followed by *query expansion* that after searching with an initial query adds to it new terms (Croft and Harper 1979). For this, *pseudo-relevance feedback (PRF)* is typically used. It scores and selects new terms according to their distribution in the feedback documents (i.e., those retrieved with the initial query), or according to the comparison of their distribution in these documents and the entire collection (Xu and Croft 2000). Re-sampling PRF terms by combining PRF results from several query sub-samples downturns the chance of adding noisy terms to the query (Collins-Thompson and Callan 2007). Twitter API terms do not allow us to run similar queries simultaneously; running them sequentially might lead to data loss at the beginning of the crisis. Hence, we cluster tweets based on which terms matched them, treating each term as a different query (Xu and Croft 2000).

Adaptive information filtering. Unlike classic query generation and expansion on static collections, the data stream relevant to crisis events evolves over time. Our query is maintained over long periods, performs a binary selection rather than compiling a ranked list of documents, and is limited in size – akin to *information filtering over streams of documents* (Allan 1996; Lanquillon and Renz 1999).

In contrast to current approaches that exploit the time dimension of a static microblog collection (Metzler, Cai, and Hovy 2012; Miyanishi, Seki, and Uehara 2013), we collect data as it is produced, rather than searching in a historical repository. Wang et al. (2013) expands a user-provided query with new hashtags to retrieve more microblog data related to given events. We automate the entire retrieval process by exploiting knowledge on past crises to generate a query, which is then expanded with terms specific to new crises.

Lexicon building. We exploit the fact of having a single domain by creating a lexicon that captures crisis-relevant terms frequently used in crises tweets, which is then adapted to a specific event (§4). Typically there are two design decisions regarding lexicons: categorize terms in a number of predefined categories (e.g., WordNet, VerbNet), and/or weight terms across one or more dimensions (e.g., SentiWordNet). The former is adopted for building broad linguistic resources with numerous dimensions. If the application domain is more focused (e.g., sentiment extraction) the later is used (Kaji and Kitsuregawa 2007), which we also adopt.

3 Datasets and Evaluation Framework

In this section we describe the input datasets we use, and the evaluation method and metrics by which we compare different alternatives.

3.1 API limits

Twitter's API for accessing tweets in real-time (the *streaming* API) has several limitations. The two that are most relevant for our work are the following.

First, tweets can be queried by content *or* by geographical location. Specifically, if both content and geographical criteria are specified, the query is interpreted as a disjunction (logical *OR*) of both. The content criterion is specified as the disjunction of up to 400 terms, in which each term is a case-insensitive conjunction of words without preserving order. The location criterion is specified as the disjunction of a set of up to 25 rectangles in coordinate space.

Second, independently of the method used to query, the resulting set is limited to 1% of the stream data. If the query matches more than 1% of the data, then the data is subsampled uniformly at random. As a result, even if we use a "blank" query (collect everything), we never obtain more than a sample of 1% of tweets. As a query becomes broader (i.e, by including more terms or a larger geographical region) at some point we start losing tweets because of this limitation. This means that "collecting everything and then post-filtering" is an ineffective sampling method: at least part of the selection must be done at query time.

Name / Type	Start / Duration	Keyword-based sampling (# of terms): Examples of terms	# of tweets	Location-based sampling Region(s)	# of tweets
Sandy Hurricane	2012-10-28 3 days	4: hurricane, hurricane sandy, frankenstorm, #sandy	2,775,812	NY City; Bergen, Ocean, Union, Atlantic, Essex, Cape May, Hudson, Middlesex & Monmouth County, NJ, US	279,454
Boston Bombings	2013-04-15 5 days	17: boston explosion, BostonMarathon, boston blast, boston terrorist, boston bomb, boston tragedy, PrayForBoston, boston attack, boston tragic	3,375,076	Suffolk and Norfolk Counties, Massachusetts, US	88,931
Oklahoma Tornado	2013-05-20 11 days	36: oklahoma tornado, oklahoma storm, oklahoma relief, oklahoma volunteer, oklahoma disaster, #moore, moore relief, moore storm, #ok, #okc	2,742,588	long. $\in [-98.25, -96.75]$ $\wedge$ lat. $\in [34.5, 35.75]$	62,237
West Texas Explosion	2013-04-17 11 days	9: #westexplosion, #westtx, west explosion, waco explosion, texas explosion, tx explosion, texas fertilizer, #prayfortexas, #prayforwest	508,333	long. $\in [-97.5, -96.5] \wedge$ lat. $\in [31.5, 32]$	16,033
Alberta Floods	2013-06-21 11 days	13: alberta flood, #abflood, canada flood, alberta flooding, alberta floods, canada flooding, canada floods, #yycflood, #yycfloods, #yycflooding	370,762	Alberta, Canada	166,012
Queensland Floods	2013-01-27 6 days	4: #qldflood, #bigwet, queensland flood, australia flood	5,393	Queensland, Australia	27,000

Table 1: Summary statistics of the six disasters and the two data samples (keyword-based and location-based).

3.2 Datasets

We use data from 6 disasters between October 2012 and July 2013, occurring in English-speaking countries (USA, Canada, and Australia) which affected up to several million people. Crisis keywords were defined by two research groups: Aron Culotta's "Data Science for Social Good" team (Ashktorab et al. 2014), and the NSF SoCS project group at Kno.e.sis using the Twitris tool (Sheth et al. 2014), who shared partial lists of tweet-ids with us. Location-based data was partially collected using Topsy analytics. As detailed in Table 1, for each disaster we use two sets of data collected from Twitter: (1) a keyword-based sample[1] and (2) a location-based sample. We note that filtering by a conjunction of keywords and locations is *not* possible using Twitter's current streaming APIs. In addition, both of these conditions occur in only a fraction of the relevant tweets (§3.3).

The *keywords-based samples* use keywords chosen by the data providers following standard practices for this type of data collection. This typically includes hashtags suggested by news media and response agencies,[2] terms that combine proper names with the canonical name of the disaster (e.g., *oklahoma tornado*), or the proper names given to meteorological phenomena (e.g., *typhoon pablo*).

The *location-based samples* are obtained by collecting all the postings containing geographical coordinates inside the affected areas. Geographical coordinates are typically added automatically by mobile devices that have a GPS sensor, in which their users have allowed this information to be attached to tweets. Location-based samples were obtained through two data providers: GNIP,[3] which allows to specify a region through a rectangle defined by geographical coordinates, or Topsy, which additionally allows to indicate the names of the places of interest (counties, states, etc.)

[1]The West Texas explosion keyword-collection was obtained from GNIP, which allows more expressive query formulation than the Twitter API. We used an estimated query that approximates this collection with a precision and recall higher than 98%.

[2]http://irevolution.net/2012/12/04/catch-22/

[3]http://www.gnip.com/

3.3 Evaluation Framework

Our filtering task can be seen as a binary classification task. The positive class corresponds to messages that are related to a crisis situation, while the negative class corresponds to the remaining messages. This is a broader, more inclusive definition than being informative (Imran et al. 2013), or enhancing situational awareness (Vieweg 2012).

Labeling crisis messages. The labeling of messages was done through the crowdsourcing platform Crowdflower[4]. For efficiency and to improve the quality of data we use to train our models, we perform a pre-filtering step. We first eliminate messages that contain less than 5 words as we deem them too short for training our lexicon. Next, we eliminate messages that are unlikely to be in English by checking that at least 66% of the words were in an English dictionary[5].

The task is designed to encourage workers to be *inclusive*, which is aligned with the goal of having high recall. We present workers a tweet and ask if it is in English and (A) directly related to a disaster, (B) indirectly related, (C) not related, or (D) not in English or not understandable. For purposes of our evaluation, the positive class is the union of tweets found to be directly and indirectly related, and the negative class is the set of tweets found to be not related.

For clarity, we include the type of disaster in the question. Example instructions appear in Figure 1. We showed 15 tweets at a time; one tweet was labeled by the authors, and used to control the quality of crowdworkers. Given the subjectivity of the task, tweets used to control quality were selected to be obvious cases.

From each crisis we labeled 10,050 tweets selected uniformly at random from the keyword-based sample (50% of labels) and location-based sample (50% of the labels). On average, about 100 workers participated in each crowd-task. We asked for 3 labels per tweet and kept the majority label. On average, 31.5% tweets were labeled as directly related, 22.2% as indirectly related, 45.8% as not-related, and 0.5% as not in English, etc.

[4]http://www.crowdflower.com/

[5]NLTK's English dictionary and the English database WordNet

Categorize tweets posted during the 2013 Oklahoma Tornado:
Read carefully the tweets and categorize them as:
A. In English and directly related to the tornado.
– "The tornado in Oklahoma was at least a mile wide"
B. In English and indirectly related to the tornado.
– "The nature power is unimaginable. Praying for all those affected."
C. In English and not related to the tornado.
– "Oklahoma played well soccer this night"
D. Not in English, too short, not readable, or other issues.
– "El tornado en Oklahoma ..."

"*Seeing everyone support #Oklahoma makes my heart smile!#oklahomatornado*"
This tweet is:
A. In English and directly related to the tornado.
B. In English and indirectly related to the tornado.
C. In English and not related to the tornado.
D. Not in English, too short, not readable, or other issues.

Figure 1: Example instructions (top) and example crowdsourcing task (bottom) used for labeling crisis messages.

Disaster	Keyword-based Prec.	Keyword-based Recall	Location-based Prec.	Location-based Recall
West Texas Explosion	98.0%	29.0%	6.7%	(100.0%)
Alberta Floods	96.0%	41.9%	8.0%	(100.0%)
Boston Bombings	86.3%	25.3%	15.9%	(100.0%)
Sandy Hurricane	92.1%	39.3%	26.1%	(100.0%)
Queensland Floods	71.2%	17.9%	8.8%	(100.0%)
Oklahoma Tornado	66.2%	45.4%	9.0%	(100.0%)
Average	85.0%	33.1%	12.4%	(100.0%)

Table 2: Precision and recall of keyword-based and location-based sampling. The task is finding crisis-related messages.

Measuring precision and recall. Evaluating *precision* is straightforward, as it corresponds to the probability that a message included in a sample belongs to the positive class. Evaluating *recall* is more difficult as it requires a complete collection containing all the crisis-related messages for each disaster. Yet, such a collection may require to label up to 300K messages to cover a single minute of Twitter activity.[6]

Since our methods rely on selecting tweets based on keywords, we evaluate them on the *location-based sample*. According to this definition, the recall of a keyword-based sampling method is the probability that a positive element in the location-based sample matches its keywords.

Table 2 evaluates the keyword-based and location-based samples using the crowdworker labels. Both precision and recall vary significantly across crises. In general, the precision of keyword-based sampling (66% to 98%) is higher than that of location-based sampling (7% to 26%). We note that the average recall of about 33% that we observe in the keyword-based samples means that about two thirds of the crisis-related messages in the location-based samples do not contain the specified keywords – that is the main motivation for the methods we describe in §4.

Further metrics. We regard the problem of collecting crisis messages as a *recall-oriented task*. Our solution should accept messages when in doubt, without accepting all messages which yields a trivial 100% recall.

There is a significant imbalance between the positive and negative classes, as seen in Table 2. So we use the metric *G-mean* – the geometric mean of the recall of the positive class and the recall of the negative class – often used to assess the classification performance on imbalanced data (Sun et al. 2007). Further, we measure the F_2 and F_1 scores, where F_k is $\frac{(1+k^2)PR}{k^2P+R}$ with P and R being precision and recall, with emphasis on the F_2 score which weights the recall more heavily for reasons we explained.

We also evaluate the proportion of different classes of messages (e.g. related to donations, warnings) in each sample. We defer the explanation of that evaluation to §5.2.

[6]https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

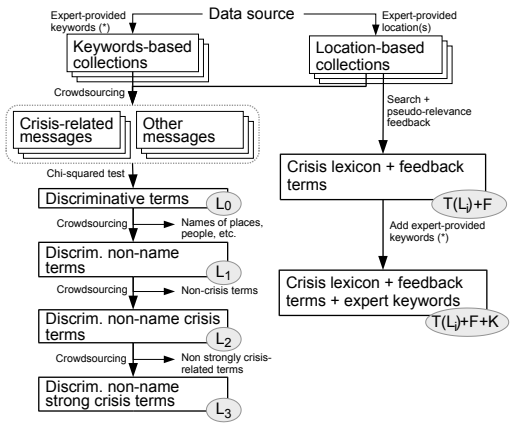


Figure 2: Steps in the lexicon construction (left), and in the evaluation of the lexicon combination with pseudo-relevance feedback and expert-provided keywords (right). $T(\cdot)$ selects the highest-scoring terms: $\mathrm{top}(\cdot)$, or the highest-scoring terms ensuring diversity: $\mathrm{topdiv}(\cdot)$

4 Proposed Method

Our method is based on creating a generic *crisis lexicon*: a list of terms to be used instead of a manual query to sample crisis-related messages. This crisis lexicon can be expanded with terms specific to a given crisis, either manually, or by using a mechanism similar to pseudo-relevance feedback.

4.1 Building the Lexicon

Figure 2 depicts the steps we take to construct the lexicon. We start by selecting the set of terms that discriminate crisis-related messages (L_0). Next, we refine this set by performing a series of curation steps filtering out both contextual and general terms as decided by crowdworkers ($L_{1...3}$). Finally, we filter out terms that frequently co-occur to maximize recall for a limited sized lexicon ($\mathrm{topdiv}(\cdot)$).

Candidate Generation Step (L_0) Term selection. Our candidate terms are word unigrams and bigrams. We start with tweets from the positive and negative classes described in §3. We remove URLs and user mentions (*@username*).

After tokenizing, we discard tokens that are too short (2 characters or less), too long (16 characters or more, typically corresponding to joined strings of words), or that correspond to punctuation, numbers, or stopwords. The remaining words are stemmed using Porter's stemmer.[7] Word unigram and bigrams are then extracted, and kept if they appear in at least 0.5% of the tweets.

Term scoring. Each term is then scored by two well-known statistical tests: chi-squared (χ^2) and point-wise mutual information (PMI), used in the past for lexicon creation (Kaji and Kitsuregawa 2007). Details are in Appendix A.

We refer to the result of a statistical test of discriminative value for a term t on a crisis c as its *discriminative score* $\mathrm{discr}(c,t)$. We rank terms according to this score, divide them in n-quantiles of one term each, and score each term t belonging to the k-th quantile according to the quantile probability ($\frac{k}{n}$). We can use this score directly, or combine it with the term's frequency in the crisis-related tweets (γ) by multiplying it with the probability of the quantile to which t belongs when the ranking is done according to γ instead of $\mathrm{discr}(c,t)$. We map scores to quantiles to give equal weight to the term's $\mathrm{discr}(c,t)$ and its frequency. The outcome is a per-crisis score of a term $s(c,t)$.

For our lexicon to be general, we look for terms that work well across a variety of crises. We tested multiple aggregations of scores across crises including median, mean, and harmonic mean. The best result was obtained when computing the mean crisis score of a term across crises, and then multiplying it by a sigmoid function to favor terms that appear in (at least 0.5% of the tweets of) several crises:

$$s_{\mathrm{agg}}(t) = \frac{1}{1+e^{-\frac{|C_t|}{2}}} \frac{1}{|C_t|} \sum_{c \in C_t} s(c,t) \qquad (1)$$

Where C_t is the set of crises in which t appears. If C_t is large enough the sigmoid function converges to 1 (> 0.9 when $|C_t| > 4$), while when the term appears to be discriminative in only one crisis, this factor is around 0.6.

Curation Steps ($L_{1...3}$). After identifying and scoring the set of candidate terms L_0, we perform a series of curation steps depicted in Figure 2 which yield increasingly filtered sets L_1 through L_3.

Removal of names (L_1). We remove terms that name contextual elements unique to a crisis. Such terms mainly fall within three categories: (a) the names of affected areas; (b) the names of individuals involved in the disaster; and (c) the names used to refer to a disaster. We ask evaluators if a term contains such proper nouns, which filtered out about 25% of the terms. The task description is in Figure 3 (top).

Removal of non-crisis terms (L_2 and L_3). Next, we filter out those words that are not specific to disasters. We consider three levels of crisis relevance: (1) *strongly crisis-specific*: the term is likely to appear more often during disasters; (2) *weakly crisis-specific*: the term *could* appear frequently during disasters; and (3) *not crisis-specific*: the term should not appear more often during disasters.

We ask evaluators to label each term with one of these categories. This task is depicted in Figure 3 (bottom). Of the terms that pass the previous filtering step (L_1), around 50% of them are filtered out by *weak* filtering (L_2) and around 65% by *strong* filtering (L_3).

[7] http://tartarus.org/~martin/PorterStemmer/

Indicate if the term is specific to a particular disaster: it contains the name of a place, the name of a person, or the name of a disaster:
A. YES, it contains a place name or it refers to the name of a region, city, etc.
– "Jersey flood"; "California people"; "okc tornado"
B. YES, it contains a person name or it refers to the name of a politician, etc.
– "Obama"; "Kevin donate"; "John hurt"
C. YES, it contains a reference to the name given to a disaster
– "Sandy hurricane"; "abfloods"; "yycfloods"
D. NO.
– "tornado"; "hurricanes"; "help rebuild"; "firefighter"; "rise"; "flame"; "every"

Indicate if the term is more likely to appear in Twitter during hazards:
A. YES, it is likely to appear more often during hazards/disasters.
– "tornado"; "donate help"; "people killed"; "state emergency"
B. NO, but could appear frequently during hazards/disasters as well.
– "power"; "water"; "nursing"; "recover"
C. NO, it shouldn't appear more often during hazards/disasters.
– "children"; "latest"; "south"; "voted"

Figure 3: Crowdtask for filtering name terms (top) and identifying strong and weak crisis-related terms (bottom).

Top-terms selection step. Twitter's API allows us to track up to $K = 400$ keywords, making this the maximum size of our lexicon. To use this allocation effectively, we test two strategies: $\mathrm{top}(\cdot)$ and $\mathrm{topdiv}(\cdot)$. The first strategy selects the top terms according to their crisis score. The second also selects the top terms according to crisis scores, but removes terms with lower crisis scores that frequently co-occur with higher score terms, as they match on a similar set of tweets. To find such a subset of terms, we compute the *independent set* on the term co-occurrence graph thresholded at a given level[8]. Given a set of queries (keywords- and location-based) and a collection of relevant tweets for each query, we build a graph G in which nodes are terms weighted by their crisis score, and between each pair of terms that co-occur in more than 50% of the tweets, we draw an unweighted edge. Then, we determine the *maximum weighted independent set (MWIS)* of G, which represents a subset of terms with high scores that rarely co-occur. Intuitively, this improves recall (since the lexicon has a limited number of terms).

The maximum independent set problem is NP-complete (Tarjan and Trojanowski 1977). We compared the approximation method in (Bar-Yehuda and Even 1985) with a simple greedy algorithm (GMWIS) that keeps the most discriminative terms that rarely co-occur. Since the latter obtains slightly higher recall scores, we present only those results obtained with GMWIS.

4.2 Applying the Lexicon

Pseudo-relevance feedback. We adapt the *generic* lexicon with terms specific to the targeted crisis. To identify such terms we employ pseudo-relevance feedback (PRF) mechanisms with the following framework:

[8] The idea of mapping terms co-occurrences on a graph is inspired by (Ntoulas, Pzerfos, and Cho 2005; Wu et al. 2006)

- Given a lexicon lex containing at most 400 terms, retrieve crisis relevant tweets in the first Δ_t hours of the event. We refer to these tweets as pseudo-relevant.
- From these tweets, extract and sort the terms (unigrams and bigrams) – which do not already belong to the lexicon – by their PRF score (explained below). Return the top k terms to be added to the lexicon.

Similar methodology has showed effectiveness in other Twitter-related search tasks (Efron et al. 2012).

PRF term scoring. PRF terms are usually scored according to their distribution in the feedback tweets, or according to the comparison of the distribution in the feedback tweets and the entire collection (Xu and Croft 2000). Due to having only the extracted PRF tweets, the scoring strategies we implement fall within the former category:

- **Frequency-based** scoring ranks PRF terms according to their frequency in the feedback tweets: $s_{prf}(t) = fr(t)$.
- **Label propagation-based** scoring propagates the scores from the query terms to PRF terms based on their co-occurrence in the feedback tweets: $s_{prf}(t) = \frac{\sum_{q \in lex} co(q,t) s_{agg}(q)}{\sum_{q \in lex} co(q,t)}$, where $co(q,t)$ is the number of co-occurrences between query term q and PRF term t, and $s_{agg}(q)$ the crisis score of q as defined in Eq. 1.

PRF term selection. To select the top PRF terms we test again the two strategies described in (§4.1): $\mathrm{top}(\cdot)$ and $\mathrm{topdiv}(\cdot)$. For $\mathrm{topdiv}(\cdot)$, we compute the MWIS based on the co-occurrence graph formed by only PRF terms.

Terms sampling. Some of the selected terms might be harmful (Cao et al. 2008). A workaround is to resample the terms based on their co-occurrence with sub-samples of the original query (Collins-Thompson and Callan 2007). The main hypotheses are that feedback documents form clusters according to the query terms that matched them, and that good PRF terms occur in multiple such clusters (Xu and Croft 2000). Yet, in contrast with Xu and Croft, we cannot make assumptions about terms distribution in the whole collection, since we only have the pseudo-relevant tweets; given the short nature of tweets we do not attempt to model their language. We use the sigmoid function to favor the PRF terms that co-occur with multiple query terms: $s_{prf}(t)/(1 + e^{-\frac{|T_{prf}(t)|}{2}})$, where $T_{prf}(t)$ is the number of terms co-occurring with term t and $fr(t)$ is t's frequency in PRF documents.

Hashtags. Hashtags are topical markers for tweets (Tsur and Rappoport 2012), used to learn about events and join the conversation (Starbird and Palen 2011). During crises, specific hashtags emerge from the start, with some quickly fading away, while others are widely adopted (Potts et al. 2011). Kamath et al. 2013 found that hashtags can reach their usage peak many hours after initial use. Thus, even if they are scarce in the beginning, if widely adopted later on, hashtags improve recall; on the other hand, if not adopted they have little impact on the retrieved data. Therefore, we lower the selection barrier for hashtags by employing a dedicated PRF-step: we add the top k hashtags (appearing in at least 3 tweets) to the query according to their frequency in the PRF documents, similar to Wang et al. 2013.

5 Experimental Evaluation

We compare against two standard practices: sampling using a manually pre-selected set of keywords, and sampling using a geographical region. The goal of the lexicon is to sample a large set of crisis-related messages; this is what we evaluate first (§5.1). Next, we see if our method introduces biases in the collection compared to existing methods (§5.2).

In both cases, we perform *cross-validation* across disasters: (1) leave one disaster dataset out; (2) build the crisis lexicon ($L_{0...3}$) using data from the remaining disasters; (3) evaluate on the excluded disaster dataset; (4) repeat the process for each of the 6 disasters, averaging the results.

5.1 Precision and Recall

We evaluate the precision and recall for sampling crisis-related messages. We also incorporate other metrics, particularly those that emphasize recall, described in §3.3.

Lexicon generation. First, we identify the best versions of our lexicon along the analyzed metrics. There are several design choices that we exhaustively explore:

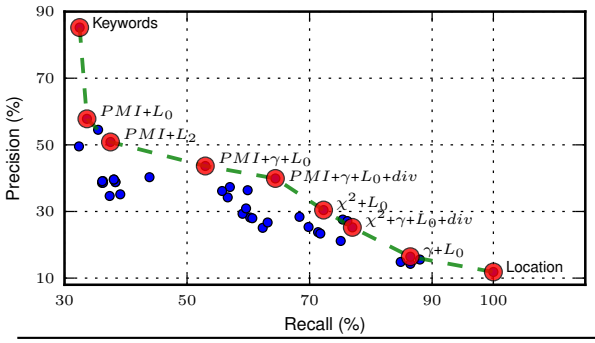
- The term scoring method (§4.1): χ^2, PMI, χ^2 + γ, PMI + γ, and γ.
- The curation steps executed (§4.1): no curation (L_0), removing names (L_1), keeping weak and strong crisis terms (L_2) and keeping strong crisis terms only (L_3).
- Whether to select the top scoring terms: $\mathrm{top}(\cdot)$, or the top scoring terms removing co-occurring terms: $\mathrm{topdiv}(\cdot)$.

This yields 40 configurations that we test along the two existing methods. Figure 4 highlights the *skyline* configurations, i.e., those for which there is no other configuration that simultaneously leads to higher recall and higher precision. Further, given that points with similar properties tend to cluster along the skyline, we keep only the points with the highest precision when they are within 5 percentage points from each other in terms of both precision and recall.

We notice that *different methods have different precision-recall trade-offs*. The term-scoring method appears to influence these trade-offs the most. Specifically, the scoring methods that penalize more a term's appearance in non-crisis tweets lead to high precision at the cost of recall (e.g., PMI); those methods that put more weight on the absolute frequency of terms in the crisis tweets lead to high recall at the cost of precision (e.g. γ). χ^2 and the combination of PMI and χ^2 with γ lead to better precision-recall trade-offs, i.e., higher F_k scores.

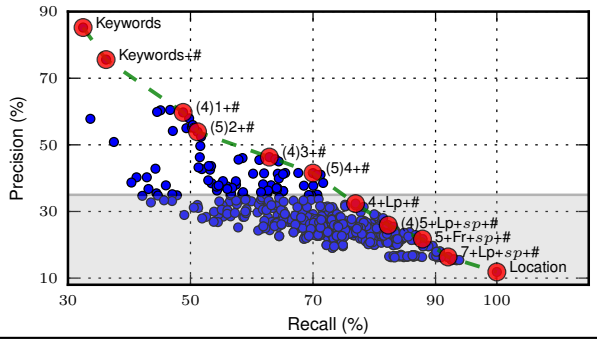
We do curation to improve precision (by removing terms that are too general) and recall (by removing terms that are too specific). Yet, curating the lexicon by removing proper nouns (L_1) lowers both the recall and precision. This effect is less pronounced when we remove terms with lower crisis scores that often co-occur with more discriminative terms ($\mathrm{topdiv}(\cdot)$). The next curation steps (L_2 and L_3) also alleviate this effect leading to higher precision overall. However, keeping only strong crisis-related terms (L_3) heavily impacts recall (the points clustered around 40% recall and precision in Figure 4).

[9] For brevity, in the rest of the paper we refer to the lexicons corresponding to these configurations by this code.



Code[9]	Config.			Prec.	Rec.	GMean	F1	F2
-	Keywords			**85.2**	32.5	56.1	45.5	36.6
-	Location			11.8	**100.0**	0.0	20.5	37.6
1	PMI	L_0	top	**57.8**	33.7	57.6	42.0	36.5
2	PMI	L_2	top	**50.9**	37.5	60.5	41.1	38.6
3	PMI+γ	L_0	top	43.7	53.0	72.7	**46.9**	50.0
4	PMI+γ	L_0	topdiv	39.9	64.4	80.2	**48.0**	56.2
5	χ^2	L_0	top	30.4	72.3	85.0	41.6	**55.0**
6	$\chi^2+\gamma$	L_0	topdiv	25.2	77.0	87.7	36.6	**52.3**
7	γ	L_0	top	16.4	**86.4**	**93.0**	26.5	43.6

Figure 4: Averaged performance of existing methods and our lexicon. Among 40 tested (small dots), the table includes the skyline configurations (large dots).



Code[9]	Config.				Prec.	Rec.	GMean	F1	F2
-	Kw.	-	#	-	75.5	36.2	59.4	47.9	40.0
p1	(4)1	-	#	-	**59.7**	48.8	69.6	**53.2**	50.4
p2	(5)2	-	#	-	53.9	51.2	70.9	50.6	50.6
p3	(4)3	-	#	-	46.3	62.9	79.2	52.3	57.8
p4	(5)4	-	#	-	41.5	**70.0**	**83.6**	50.6	**60.0**
-	4	Lp	#	$\neg$ sp	32.4	76.9	87.6	43.1	56.8
-	(4)5	-	#	-	25.9	82.3	90.7	37.9	54.5
-	5	Fr	#	sp	21.7	87.9	93.7	33.2	51.0
-	7	Lp	#	sp	16.4	95.9	95.7	26.6	44.5

Figure 5: Averaged performance of existing methods and our lexicon with PRF. From about 700 tested (small dots), the table includes the skyline configurations (large dots). The gray area marks the configurations with precision below 35% and places the corresponding skyline points at the end of the table. $(L_i)L_j$ means that we run PRF with L_i and then add the PRF terms to L_j, where L_i is a lexicon code from Figure 4.

Lexicon expansion. With the parameter combinations from Figure 4 (7 options), we test the performance of our lexicon when using various pseudo-relevance feedback (PRF) mechanisms (§4.2). We explore the following design choices:

- PRF term scoring (§4.2): frequency (Fr) and label propagation (Lp).
- Whether to select the top scoring terms: $\mathrm{top}(\cdot)$, or the top scoring terms removing co-occurring terms: $\mathrm{topdiv}(\cdot)$.
- Whether to favor terms that co-occur with more query terms (§4.2): sp, or not: $\neg\,\mathrm{sp}$.
- Whether to use *only* a hashtag (#) dedicated PRF, combine it with the PRF for terms (as defined by the previous choices), or use the later alone (§4.2).

We also combine lexicons by first running PRF with L_i, select the PRF terms, and then add them to L_j, where L_i, L_j are lexicons obtained with the skyline configurations of Figure 4; combination denoted $(L_i)L_j$. This yields about 700 configurations to test. For these tests we set the number of PRF terms to 30, and PRF interval to $\Delta_t = 3$ hours. We assume the data collection, and the PRF, start simultaneously with the keywords-based collection. Results are in Figure 5.

We notice that PRF boosts recall, but has little impact on precision. Further, the lexicon combinations with the #-dedicated PRF lead to better precision-recall trade-offs when L_i has high recall and L_j has high precision.

Expert-defined terms. To analyze how the expert-defined crisis-specific terms and the lexicon complement each other, we add the former to the queries corresponding to the top skyline configurations depicted in Figure 5.

As shown on Table 3, such combinations generally lead to improvements over both the keywords and the lexicon (e.g., up to 40 percentage points recall over the crisis-specific keywords). The only metric we do not improve on is the *precision* of the keyword collection, yet this is an upper bound for precision as the expert-edited keywords are chosen to be specific only to a given disaster. Further, though the precision decreases, the combination leads to better precision-recall trade-offs, as it improves over the F-score metrics. p2 leads to the highest gains over the lexicon-based approach and over the F1-score of the keyword-based approach; meaning that the samples obtained with p2 and those obtain with the crisis-specific keywords overlap the least.

Performance over time. Finally, to analyze the performance variation over time, we test two design decisions: running PRF only one time at the beginning of the crisis (one-time PRF), or re-running PRF after every 24 hours (online PRF). We measure the average performance's variation across the first three days from the start of the keyword-collections.[10] Figure 8 shows the performance of the lexicon with both one-time PRF and online PRF in terms of recall and F1-score relative to the crisis-specific keywords, which is the reference values. We omit the corresponding precision plots, but note that an increase in recall with no improvement in F1-score

[10]We restrict this analysis to the first three days for two reasons: all collections span across at least three days, and, typically, the largest volumes of tweets happen in the first days of the event.

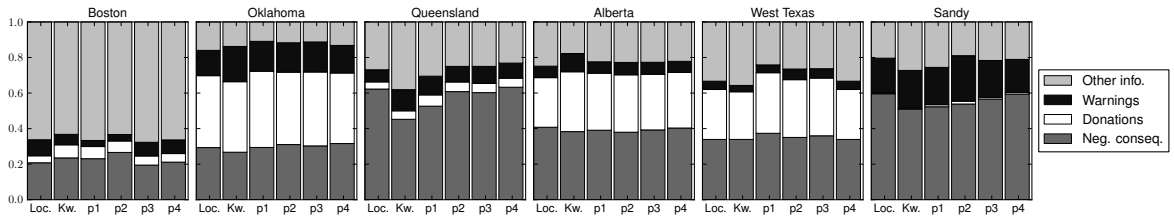


Figure 6: Tweet distribution per type of information for each sampling method. The average BC coefficient between the distribution of the message types in the location-based collection and in data sampled from it with our lexicon and the keywords: $BC(keywords) = 0.994$, $BC(p1) = 0.995$, $BC(p2) = 0.996$, $BC(p3) = 0.998$, $BC(p4) = 0.999$

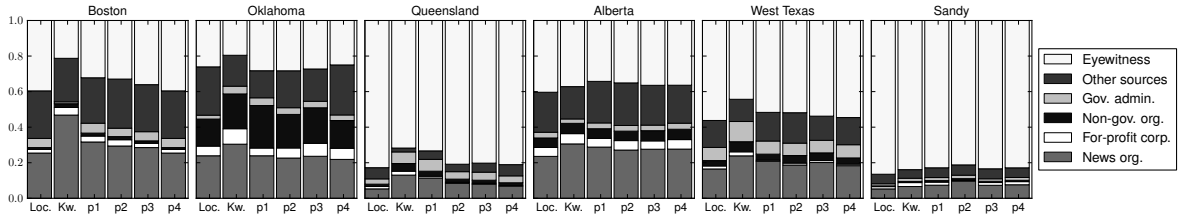
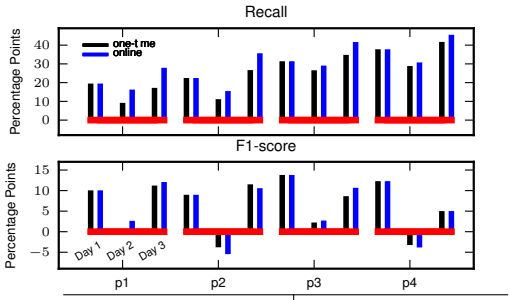


Figure 7: Tweet distribution per type of source for each sampling method. The average BC coefficient between the distribution of the message source in the location-based collection and in data sampled from it with our lexicon and the keywords: $BC(keywords) = 0.984$, $BC(p1) = 0.993$, $BC(p2) = 0.996$, $BC(p3) = 0.997$, $BC(p4) = 0.999$

indicates a loss in precision.

In our experiments, the lexicon based approaches do better on average (in the range of 20 to 40 percentage points for recall and 9 to 13 percentage points for F1-score) towards the beginning of the crisis compared to the crisis specific keywords. Then we see a drop in the performance relative to the keywords which might be due to more users conforming to keywords use as the event gets global coverage, followed by an increase when the event loses coverage. Finally, although employing online PRF leads to better recall values later on in the crisis, it's improvement in terms of F1-score over one-time PRF is only marginal.



	Recall			F1-score		
	Day 1	Day 2	Day 3	Day 1	Day 2	Day 3
Keywords	35.3	**39.0**	31.6	49.4	**53.4**	41.7

Figure 8: Relative performance over time of our lexicon with one-time PRF and online PRF re: crisis-specific keywords. The table contains the reference performance by the keywords – represented by the (red) horizontal line.

5.2 Distribution of message types

We measure changes in the distribution of tweets of different types, as sampling by keywords may introduce *biases* that favor one class of tweets at the expense of another. We evaluate by asking crowdworkers to categorize tweets, and then measure the divergence between the distribution of tweets into categories across the sampling methods. We repeat this twice using three categorizations: informativeness, information type and information source (details in Appendix B).

First we check if any sampling method biases the collection towards the tweets deemed informative by crowdworkers. With one exception, we find only marginal differences across crises; looking at crisis-relevant tweets, we find that between the lexicon and the crisis-specific keywords there is a difference of less than 10 percentage points regarding the proportion of informative tweets. The (reference) location-based samples have lower proportions of informative tweets than the lexicon and keywords-based samples. The exception is Hurricane Sandy, for which the p2 configuration collects more informative tweets (about 18 percentage points) than the keywords sample.

Figures 6 and 7 depict the tweets distribution according to the type and source of information. We also show the Bhattacharyya coefficient (BC) which quantifies the overlap between the reference location-based collection and lexicon and keyword-based samples in terms of information type and source; high values indicate high similarity.

We notice large variations in tweet distributions according to both the information type and source across crises; yet it has little to no impact on the sampling methods' ability to preserve the distributions. Generally, high-precision methods diverge more from the reference sample, with the keywords being the least representative, e.g., it collects more tweets coming from news organizations and fewer eyewitness re-

ports (Figure 7). In contrast, our lexicon better preserves the reference distribution, with a BC close to 1.

6 Conclusions

We have described a methodology for constructing an effective, general lexicon for monitoring crisis events. Our experiments demonstrate a range of precision and recall operating points previously not well understood when using only keyword or location-based sampling. This work provides researchers an informed strategy for assembling a set of relevant tweets. This is a fundamental technology for automatic linguistic analysis tools such as temporal summarization.

The impact of these results goes beyond an algorithmic understanding. We show that the amount of data that it is currently mined represents only a fraction of the data posted during disasters. We believe that such lexicons can support others interested in increasing recall, but who may not have the ability to finely tune their lexicons.

There are many directions in which to take this work. First, users are often interested in classifications more finely grained than 'relevant' or 'nonrelevant': e.g., emergency responders may be interested in personal or property loss tweets, each of which will admit its own lexicon. Second, though our techniques are in principle language-independent and domain-independent, we want to build lexicons which demonstrate this. Third, when using a lexicon to collect data through an API, if the API is more limited or less limited, or limited on a different way, our results may have to be adapted. Fourth, we would like to keep human effort to a minimum –mostly because we may want to build a specialized lexicon in a short time– and we are working on methods to simplify the manual steps of the process.

Reproducibility. The crisis lexicon, the list of keywords, geographical regions, etc. along with the labeled datasets as sets of (tweet-ids, label, and metadata) are available for research purposes at http://crisislex.org/.

Acknowledgments. Work done while Alexandra Olteanu was doing an internship at QCRI. We are grateful to Topsy for providing data corresponding to four geo-collections we analyzed, and to Hemant Purohit and Aron Culotta for sharing with us the tweet ids for the rest of the collections.

Config.	Prec.	Rec.	Gmean	F1	F2
p1	60.8 (-24.4/1.1)	55.7 (23.1/6.9)	74.2 (18.2/4.5)	56.1 (11.7/4.1)	57.3 (19.5/5.6)
p2	56.9 (-28.3/**3.1**)	60.7 (28.4/**8.4**)	77.7 (21.7/**6.0**)	**57.7** (**12.2**/**6.8**)	59.2 (22.7/**7.8**)
p3	47.7 (-37.4/1.6)	66.6 (34.1/3.7)	81.5 (25.5/2.2)	54.8 (9.3/2.7)	61.0 (24.5/3.3)
p4	42.3 (-42.8/1.0)	**73.5** (**41.0**/3.5)	**85.7** (**29.6**/1.8)	52.4 (6.9/2.3)	**62.7** (**26.2**/2.7)

Table 3: Average performance of our lexicon when combined with crisis-specific keywords. We also report (the improvement over such keywords/the improvement over the method without these keywords) as percentage points.

A Statistical tests for terms

For each term t we compute the following contingency table:

	related	not related
t	$n(\mathrm{t}, \mathrm{rel})$	$n(\mathrm{t}, \neg\,\mathrm{rel})$
$\neg$ t	$n(\neg\,\mathrm{t}, \mathrm{rel})$	$n(\neg\,\mathrm{t}, \neg\,\mathrm{rel})$

where $n(t, c)$ is the number of tweets belonging to class c in which term t appears, $n(\neg\,\mathrm{t}, \mathrm{rel})$ the number of tweets in which term t does not appear and $c \in \{\mathrm{rel}, \neg\,\mathrm{rel}\}$. Then, similarly with (Kaji and Kitsuregawa 2007), we use two popular statistical measures to estimate how strong the association between a term and the crisis-related tweets is (the *discriminative score*): Chi-square (χ^2) and Pointwise Mutual Information (PMI).

χ^2-based crisis score. The statistical measure χ^2 tests whether a term t occurrence is independent of the tweet being about a disaster or not; and is defined as follows:

$$\chi^2 = \sum_{x \in \{\mathrm{t}, \neg\,\mathrm{t}\}} \sum_{c \in \{\mathrm{rel}, \neg\,\mathrm{rel}\}} \frac{(n(x,c) - E[n(x,c)])^2}{E[n(x,c)]}$$

where $E[n(x, c)]$ is the expected value for $n(x, c)$.

Although χ^2 estimates the discriminative power of a term t towards one of the classes, it does not indicate if t is discriminative for the crisis-related tweets. So we ignore the χ^2 when t appears more often in the non-crisis-related tweets and define the crisis score as follows:

$$cs_{\chi^2}(t) = \begin{cases} \chi^2 & \text{if } n(\mathrm{t}, \mathrm{rel}) > n(\mathrm{t}, \neg\,\mathrm{rel}) \\ 0 & \text{otherwise} \end{cases}$$

PMI-based crisis score. PMI measure the relatedness between term t and a certain class c and it is defined as (Church and Hanks 1990):

$$\mathrm{PMI}(t, c) = \log_2 \frac{P(t, c)}{P(t)P(c)}$$

where $P(t, c)$ is the joint probability of t and c, and $P(t)$ and $P(c)$ are the marginal probability of t and c.

Even if PMI measures how strongly associated term t and class c are, it does not account for how strongly associated t and the other class are. So we compute the crisis score as the difference between the association strength with crisis-related tweets and the association strength with non-crisis-related tweets (Kaji and Kitsuregawa 2007):

$$cs_{\mathrm{PMI}}(\mathrm{t}) = \mathrm{PMI}(\mathrm{t}, \mathrm{rel}) - \mathrm{PMI}(\mathrm{t}, \neg\,\mathrm{rel}) = \log_2 \frac{p(\mathrm{t}\,|\,\mathrm{rel})}{p(\mathrm{t}\,|\,\neg\,\mathrm{rel})}$$

where $p(\mathrm{t}\,|\,\mathrm{rel})$ and $p(\mathrm{t}\,|\,\neg\,\mathrm{rel})$ are the probabilities of t to appear in crisis-related, respectively non-crisis-related tweets:

$$p(\mathrm{t}\,|\,\mathrm{rel}) = \frac{n(\mathrm{t}, \mathrm{rel})}{n(\mathrm{t}, \mathrm{rel}) + n(\neg\,\mathrm{t}, \mathrm{rel})}$$

$$p(\mathrm{t}\,|\,\neg\,\mathrm{rel}) = \frac{n(\mathrm{t}, \neg\,\mathrm{rel})}{n(\mathrm{t}, \neg\,\mathrm{rel}) + n(\neg\,\mathrm{t}, \neg\,\mathrm{rel})}$$

This yields positive scores when t has a higher probability of appearing in crisis tweets than in non-crisis tweets, and negative otherwise. Therefore, we consider only positive values.

B Message Types Categorization

We label crisis-relevant tweets distribution along two main categorizations: information type, and information source.

For each, we present workers a tweet and ask them to label it with the likeliest category (see Figure 9). For quality control, one of every 10 tweets presented to a worker was labeled by one of the authors and was chosen to be an obvious case.

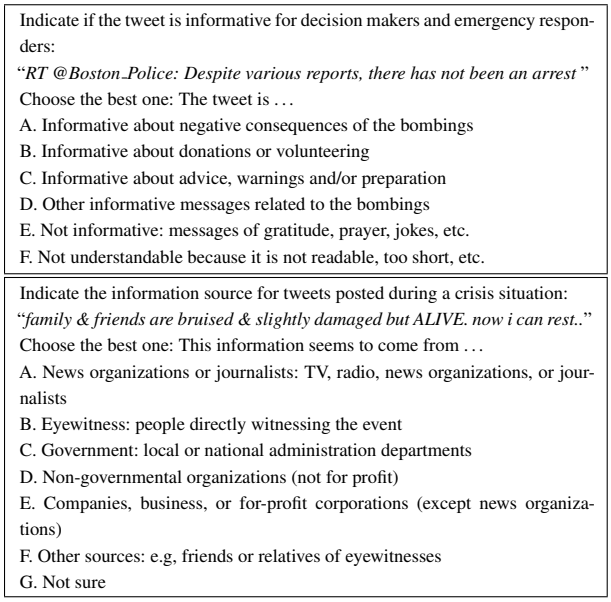


Figure 9: Crowd-tasks for categorizing tweets according to informativeness and type (top), and source (bottom).

References

Allan, J. 1996. Incremental relevance feedback for information filtering. In *SIGIR*.

Arguello, J.; Diaz, F.; and Paiement, J.-F. 2010. Vertical selection in the presence of unlabeled verticals. In *SIGIR*.

Ashktorab, Z.; Brown, C.; Nandi, M.; and Culotta, A. 2014. Tweedr: Mining twitter to inform disaster response. *ISCRAM*.

Bar-Yehuda, R., and Even, S. 1985. A local-ratio theorem for approximating the weighted vertex cover problem. *Annals of Discrete Mathematics*.

Bruns, A., and Liang, Y. E. 2012. Tools and methods for capturing twitter data during natural disasters. *First Monday* 17(4).

Bruns, A.; Burgess, J. E.; Crawford, K.; and Shaw, F. 2012. # qldfloods and@ qpsmedia: Crisis communication on twitter in the 2011 south east queensland floods. *ARC Centre, Queensland University of Technology*.

Cao, G.; Nie, J.-Y.; Gao, J.; and Robertson, S. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR*.

Church, K. W., and Hanks, P. 1990. Word association norms, mutual information, and lexicography. *Comp. linguistics*.

Collins-Thompson, K., and Callan, J. 2007. Estimation and use of uncertainty in pseudo-relevance feedback. In *SIGIR*.

Croft, W. B., and Harper, D. J. 1979. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation* 35(4).

Efron, M.; Deisner, J.; Organisciak, P.; Sherman, G.; and Lucic, A. 2012. The university of illinois graduate school of library and information science at trec 2012. In *TREC*.

Hughes, A. L., and Palen, L. 2009. Twitter adoption and use in mass convergence and emergency events. *ISCRAM*.

Imran, M.; Elbassuoni, S.; Castillo, C.; Diaz, F.; and Meier, P. 2013. Practical extraction of disaster-relevant information from social media. In *WWW Companion*.

Kaji, N., and Kitsuregawa, M. 2007. Building lexicon for sentiment analysis from massive collection of html documents. In *EMNLP-CoNLL*.

Kamath, K. Y.; Caverlee, J.; Lee, K.; and Cheng, Z. 2013. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *WWW*.

Lanquillon, C., and Renz, I. 1999. Adaptive information filtering: detecting changes in text streams. In *CIKM*.

Metzler, D.; Cai, C.; and Hovy, E. 2012. Structured event retrieval over microblog archives. In *NAACL HLT*.

Miyanishi, T.; Seki, K.; and Uehara, K. 2013. Improving pseudo-relevance feedback via tweet selection. In *CIKM*.

Ntoulas, A.; Pzerfos, P.; and Cho, J. 2005. Downloading textual hidden web content through keyword queries. In *JCDL*.

Potts, L.; Seitzinger, J.; Jones, D.; and Harrison, A. 2011. Tweeting disaster: Hashtag constructions and collisions. In *SIGDOC*.

Qu, Y.; Huang, C.; Zhang, P.; and Zhang, J. 2011. Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *CSCW*.

Sheth, A.; Jadhav, A.; Kapanipathi, P.; Lu, C.; Purohit, H.; Smith, A. G.; and Wang, W. 2014. Chapter title: Twitris-a system for collective social intelligence. *Encyclopedia of Social Network Analysis and Mining*.

Starbird, K., and Palen, L. 2010. Pass it on?: Retweeting in mass emergency. In *ISCRAM*.

Starbird, K., and Palen, L. 2011. "voluntweeters": Self-organizing by digital volunteers in times of crisis. In *CHI*.

Sun, Y.; Kamel, M. S.; Wong, A. K.; and Wang, Y. 2007. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*.

Tarjan, R. E., and Trojanowski, A. E. 1977. Finding a maximum independent set. *SIAM Journal on Computing*.

Tsur, O., and Rappoport, A. 2012. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *WSDM*.

Vieweg, S.; Hughes, A. L.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *CHI*.

Vieweg, S. 2012. *Situational Awareness in Mass Emergency: A Behavioral and Linguistic Analysis of Microblogged Communications*. Ph.D. Dissertation, University of Colorado at Boulder.

Wang, X.; Tokarchuk, L.; Cuadrado, F.; and Poslad, S. 2013. Exploiting hashtags for adaptive microblog crawling. In *ASONAM*.

Wu, P.; Wen, J.-R.; Liu, H.; and Ma, W.-Y. 2006. Query selection techniques for efficient crawling of structured web sources. In *ICDE*.

Xu, J., and Croft, W. B. 2000. Improving the effectiveness of information retrieval with local context analysis. *ACM TOIS*.

Yin, J.; Lampert, A.; Cameron, M.; Robinson, B.; and Power, R. 2012. Using social media to enhance emergency situation awareness. *IEEE Intelligent Systems*.