Exploring the dangers of searching Twitter using pre-defined keywords

ABSTRACT

Several studies on Twitter usage during disasters analyze tweets collected using keywords pre-defined by researchers. While recent efforts have worked to improve this methodology, open questions remain about which keywords can be used to uncover tweets contributing to situational awareness (SA) and the quality of tweets returned using different terms.

Herein we explore the varying extent to which a large set of keywords, extracted from Ushahidi reports, can be used to find tweets providing SA in the week following the 2010 Haitian earthquake and how the tweets collected differ from those used in previous work. Using a novel methodology, we quantify the benefit of different search terms, finding that specific locations in Haiti were the best search terms for finding tweets containing SA information. We also find that certain terms that could have been useful in finding tweets containing situational awareness never co-occurred with terms used to search Twitter in previous works, suggesting difficulties in uncovering them from the Twitter-sphere itself. Finally, we observe that tweets providing SA are sparser in samples that didn't contain keywords utilized in previous work on the earthquake.

Keywords

Social Media and Disasters, Twitter, Sampling from Twitter, Data mining

INTRODUCTION

Researchers studying Twitter usage patterns both generally (boyd and Crawford, 2012) and during disasters (Munro, 2011; Oh et al., 2010; Starbird et al., 2012, 2010) frequently investigate tweets collected using a single, static set of pre-defined keywords. For example, six out of the eight articles studying Twitter in last year's ISCRAM Proceedings used some form of a static, pre-defined keyword search. Though researchers do not assume that the resulting tweets represent the complete set of messages relevant to their particular topic, results are nonetheless constrained to unique datasets based on different search terms. The extent to which any of these datasets capture the full set of disaster-related tweets is largely unknown.

Volunteer aid workers during disasters have also used keyword-based searches of Twitter in attempts to increase *situational awareness (SA)* (Morrow et al., 2011). Quoting Sarter and Woods, 1991, Vieweg et al. (2010) define SA as "all knowledge that is accessible and can be integrated into a coherent picture, when required, to assess and cope with a situation". While practical, the use of pre-defined keyword to obtain tweet from the Twitter API(s) presents at least two difficulties to the searcher. First, predefined keywords cannot adapt to the disaster. No one knows in advance which cities, streets or buildings, and consequently which terms, will play a key role at different times during and after a disaster. Similarly, Potts et al. (2011) find that knowing which hashtags will become popular in a disaster setting is difficult, leading to problems in capturing relevant tweets particularly in a given disaster's early stages. Second, researchers have only recently begun to understand the lexicon that differentiates a tweet which provides SA from one that does not (Verma et al., 2011; Vieweg, 2012). Such work provides little indication that tweets providing SA always use terms that would be intuitive to search for, and in any case is completed on tweets which already contain at least one such term anyways.

In recognizing these issues, recent work both within the disaster response community (Abel et al., 2012) and outside of it (Li et al., 2013; Lin et al., 2011; Tsagkias et al., 2011) has suggested approaches to automatically selecting the keywords used to search Twitter for tweets relevant to a particular topic. Exploiting ideas that rely on correlations between relevant tweets and keywords as opposed to researcher intuition, these approaches leverage keywords one might not initially think to search for to find relevant tweets. For example, Li et al. (2013) observe that the word "heroin" was one of the best terms for finding tweets relevant to "any crime or disaster". Though this fits post-hoc intuition, one likely would not select this term initially. This suggests that statistical approaches to keyword selection are able to utilize "unintuitive" keywords to find relevant tweets.

While such work has presented important new methods for adapting keyword searches, the following three research questions with respect to keyword searching and Twitter during disasters still remain largely unaddressed:

> *RQ1:* Can we characterize the most useful "unintuitive" keywords for capturing tweets containing SA?
>
> *RQ2:* What is the overlap between samples of tweets generated from searching with intuitive, pre-defined keywords and those generated using unintuitive terms to search?
>
> *RQ3:* What is the quality of tweets that contain unintuitive keywords that are missed by searches using intuitive terms?

In the present work, we use 81M tweets (approximately 15% of all tweets during the time period) from the week following the 2010 Haitian earthquake to explore these three questions. For the purposes of this case study, we define *intuitive* keywords as those used by authors of all previous works we are aware of to capture tweets on the Haitian disaster. We define *unintuitive* keywords as those relevant to the disaster not used by previous work[1]. To capture these unintuitive terms, we use a novel approach of pulling terms from another crowd-sourced dataset available at the time, namely detailed reports submitted by volunteers running an instance of the crowd-sourced crisis mapping tool Ushahidi[2]. We take from these reports locations, named entities and actions using both standard IR techniques and heuristic rules and then vet the resulting list by hand.

We use our sets of intuitive and unintuitive terms to answer our three research questions. With respect to *RQ1*, we provide a novel methodology to rank the relevance to the disaster of each unintuitive term. Our approach is based on the idea of assessing the expected *benefit* of tracking (searching for) a particular term in the aftermath of the disaster. We find that the unintuitive terms expected to be most useful in searching for important disaster-related tweets are locations specific to the disaster affected area, a claim suggested by previous work (Abel et al., 2012; Vieweg et al., 2010). We present several empirical examples confirming this point and, through our proposed metric and the usage of an alternative avenue of crowd-sourced data, provide guidelines for efficiently obtaining these terms locations.

With respect to *RQ2*, we find that tweets captured using intuitive keywords rarely contain unintuitive keywords as well. This suggests that approaches which pull keywords from outside the Twitter-sphere may be able to capture a larger portion of relevant tweets. Finally, we develop a heuristic to classify tweets as providing situational awareness or not and use this classifier to address *RQ3.* We use the process described in (Vieweg, 2012) to determine the quality of the tweets our classifier uncovers relative to a similar set obtained using intuitive keywords. We find that samples without intuitive keywords provide approximately half as many tweets that actually provide situational awareness as the tweets studied by Vieweg (2012). While our results are gathered using different classifiers, the resulting empirical values and qualitative observations of the data suggest that situationally aware tweets may be sparser in tweets without intuitive keywords.

Our work contradicts claims that sampling from big data is inherently biased and thus unusable in disaster settings. Rather, we find that even pre-defined keyword sampling appears to be "good enough", and that new methods, including the one proposed here, hold promise in further reaching important information that can genuinely impact the situational awareness of decision makers and first responders. We also show that for practitioners who cannot afford the luxury of one of these systems, including more relevant locations in their keyword searches should help capture more tweets containing situational awareness information.

[1] As we will see, these unintuitive keywords often make intuitive sense, but the distinction here is made in an attempt to provide the most unbiased approach we could take to defining "intuitive" in this scope.

[2] http://www.ushahidi.com

Field Name	Description
Date	The date and time at which the incident was logged by an Ushahidi volunteer
Title	A volunteer-specified title for the report
Location	A volunteer-specified location
Content	A write-up of the reported incident, usually including direct message content

Table 1 Ushahidi report fields

RELATED WORK

Keyword sampling on Twitter

While pre-defined keyword sampling appears to be the stock approach to capturing Twitter data, it is far from the only one (Landwehr and Carley, 2014). One increasingly popular method is to sample by first finding users who are providing SA information and then pulling all the tweets from those users (Sarcevic et al., 2012; Starbird et al., 2012, 2010). Another approach is to use geospatial bounding boxes to obtain tweets within the area affected by the disaster (Kumar et al., 2011; MacEachren et al., 2011). While such methods appear to provide highly relevant results, keyword-based searches are most likely to incorporate both users who provide only a minute number of SA tweets and those who do not geo-tag their tweets. Thus, we focus here only on keyword-based searches.

Abel et al. (2012) design a system that takes named entities (including locations) from both official reports on disasters and previously captured tweets. They show that doing so provides a more relevant sample than simply using pre-defined keywords on a variety of datasets. Li et al. (2013) develop a more general system that works with any classifier to iteratively find tweets most relevant to a particular topic as time progresses. In principle, their system could be combined with existing classifiers for situational awareness (Corvey et al., 2012; Munro, 2011; Verma et al., 2011) to dynamically select keywords. Our work is in a similar vein to both of these articles. However, instead of drawing keywords from official reports or from the Twitter stream itself, we utilize a different set of crowd-sourced data. We also focus much more heavily on a qualitative analysis of the keywords used for search rather than the quality of the resulting dataset.

Ushahidi

Within a few hours of the 2010 Haiti earthquake, volunteer crisis workers created a deployment of the crowd-sourced mapping tool Ushahidi and began combing Twitter, media sources and organization websites for information on the situation (Morrow et al., 2011). Starting days after the earthquake, volunteers also began receiving SMS messages at a rate of 1,000-2,000 per day from Mission 4636, an automated system set up to collect text messages from people within Haiti. Ushahidi volunteers used the information gathered to create a dynamic map of the Haiti crisis, inform the United States Coast Guard of particular situations and to file reports on actionable information items which were then logged electronically (Munro, 2011). It is these reports, made publicly available for study, which we use here. Each report in the data we obtained consists of seven fields- the four used in the present work are described in Table 1. Of particular interest is the content section, from which each of the other fields may be- and often were- derived by Ushahidi volunteers.

Studies on Ushahidi reports after the Haitian disaster revealed issues with their ability to provide actionable information to aid workers (Munro, 2013). For example, a Harvard Humanitarian Initiative report found that as many as 70% of the Ushahidi reports consisted of requests for the rescue of friends and family whom the requestor already knew to be dead (Harvard Humanitarian Initiative, 2011). While not in a position to confirm or rebut these claims, we do believe that regardless of their usefulness for first responders, Ushahidi reports provide a distinct picture of the concerns and language being used by victims and volunteer aid workers during the disaster. This language can be used to discover additional useful information.

METHODOLOGY

Data

We use a corpus of tweets obtained from the authors of (O'Connor et al., 2010) that were released for research for this specific work. It is a random collection of approximately 90M mostly English tweets pulled from the gardenhose from January 7th through January 20th of 2010 (81M tweets were sent after the earthquake). At the

Work	Keywords	Sample Size
(Munro and Manning, 2012; Munro, 2011)	#haiti	~40,000
(Oh et al., 2010)	#haitiearthquake	962
(Corvey et al., 2012; Verma et al., 2011; Vieweg, 2012)	haiti, earthquake, quake, shaking, tsunami, ouest, port-au-prince, tremblement, tremblement de terre	4M (230,000)
(Sarcevic et al., 2012)	earthquake, port-au-prince, ouest, tsunami, haiti, tremblement	3.28M (~16,000)

Table 2 Known articles on the Haitian Earthquake, the search terms used, and the size of the resulting dataset. Sizes in parentheses represent sizes of final corpora after additional sampling and subsampling techniques were applied

time the tweets were drawn, the gardenhose returned about 15% of all public tweets.

From the Ushahidi reports described above, we use the 1105 reports which were filed during the week after the earthquake, matching the period of our Twitter dataset.

Methods

Our methodology proceeded in a series of steps which incorporated selecting intuitive terms from previous work, extracting unintuitive terms from the Ushahidi reports, scoring these unintuitive terms for their relevance to the earthquake and then using these scores to, in turn, provide a quality score for each tweet. All code for the present work, as well as all publically available data and all details of data cleaning, can be found at REMOVED_FOR _REVIEW.

To determine the set of intuitive keywords, we draw all keywords used in seven studies of the Haitian disaster and Twitter. Table 2 shows these articles, the search terms used and the resulting size of each study's dataset. Note that in some cases, these "keywords" really consisted of multiple words. In keeping with previous work on Twitter, we will assume that "keywords" represent phrases (e.g. "tremblement de terre") that must match in their entirety to be considered "found" in a particular tweet.

We extract three different types of unintuitive keywords from the Ushahidi reports: entities, actions and locations. In order to do so, we use a combination of open-source text mining tools and heuristics developed to take advantage of the structure of the Ushahidi reports. The natural language toolkit (*nltk)* is an open-source text mining tool. We use *nltk* to capture actions and named entities from the content field of each report. Actions are drawn by invoking *nltk*'s default part of speech (POS) tagger and capturing any term in the text classified as a verb. Named entities were drawn using *nltk*'s default Named Entity Recognizer (NER). While the NER had the capability to distinguish between actors and locations, we found that the categorizations it provided were rarely accurate for this particular application area. Because of this, we manually distinguished between locations (e.g. "Port au Prince" and "Texaco Station"), and entities, which represented all named entities discovered which we could not classify as a location.

While the Ushahidi reports' content fields held unstructured text, the location and title fields both held highly structured data. The location field specified places in a hierarchical manner, split by commas (e.g. "Texaco Station, Port-au-Prince, Haiti"). We split the location field by commas and used each term as a unique location. The title field held a concise summary of the content of the report. We pull both locations and entities from the title by first extracting series of words beginning with capitalized letters and combining them into a single keyword. Any keyword following any of the words "at", "in", "on" or "by" was categorized as a location. Otherwise, it was considered an entity. As suggested, each Ushaidi report could therefore be associated with multiple locations, entities and actions across all three fields. While future work may leverage distinctions across fields (e.g. Tsagkias et al., 2011) and attempt to uncover keywords that best define each report, the present work simply combines all terms found in any field into a single collection of terms.

After automatically collecting locations, entities and actions, we manually filter for keywords that do not represent these types of terms. For actions, this meant removing terms mis-categorized by the POS tagger and removing Creole terms which did not translate to verbs according to Google Translate. For entities, we removed

Type	Before Cleaning	After Cleaning
Actions	933	708 (75.8% of original)
Entities	2229	1651 (74.1% of original)
Locations	842	741 (88.0% of original)

Table 3 Number of keywords for each type before and after manual cleaning.

nonsense terms and terms that could never refer to a specific collection of entities (e.g. "everyone", but not "someone"). For locations, we removed those which obviously referred to places outside of Haiti, as well as nonsense terms.

Table 3 provides the number of locations, actions and entities before and after cleaning. Having created the cleaned list of locations, actions and entities, we needed to distinguish the importance of each remaining term in searching for disaster-related tweets, particularly tweets that provided SA. Given a representation of this importance, we would be able to assess both the usefulness of each term ($RQ1$) and the quality of tweets that contained them ($RQ3$). Qualitative analysis suggested five general *categories* that keywords fell into (regardless of type). First, there were terms like "Jacmel" (a city in Haiti), which were expected to provide a large number of relevant tweets and very few tweets unrelated to the disaster. Second were terms like "trapped", which were likely to produce tweets relevant to the disaster but also were likely to produce some noise because of the variety of contexts they could be used in. Third were terms such as "food", which were expected to product a significant amount of noise but that may also have been relevant in a select set of SA tweets. Fourth were highly specific terms found in only one of the Ushahidi reports, like street addresses, which would likely prove important if found but were unlikely to be observed in any tweets. Finally were terms found in only a single Ushahidi report, such as (possibly mis-spelled) Haitian first names, which even if found were unlikely to produce information relevant to SA.

In order to capture these different categories of tweets quantitatively, we score terms based both on their relevance to the disaster, as determined by their frequency of use in the Ushahidi reports, and on their relevance outside the context of the disaster, as determined by their frequency of use in tweets we know were not related. The frequency of use in Ushahidi reports was determined by counting the number of reports in which each term appeared one or more times. This produced the set U, where u_k refers to the score for a specific keyword k. To determine the frequency of use in tweets not relevant to the disaster, we calculated the average number of tweets each term appeared in across nine independent sets of 1M tweets that were sent *before the earthquake occurred.* This produced the set T, where t_k refers to the score for a specific keyword k. After obtaining U and T, we take the logarithm of all values. For terms where $t_k = 0$ (and thus the logarithm was undefined), we set $t_k = \min(\{t_x | \ x \ is \ found \ at \ least \ once\}) - 1$, thus scoring them an order of magnitude lower than any other term. Following this log-scaling, we then further scale each set to have a standard deviation of 1 and a mean of $\min(U) + 1$ and $\min(T) + 1$, respectively (to ensure there are no values less than 1). The expected benefit score of each keyword is then calculated as $benefit\ score(k) = \frac{u_k}{t_k}$. Note that for terms that never appeared in the pre-disaster tweets and appeared in only one Ushahidi report, the benefit score equaled one. Such terms, which fell into the fourth and fifth qualitative categories discussed above, were thus given a "medium" weight to represent our uncertainty as to whether they were noise or not.

The final step in our methodology was to provide some mechanism to determine the quality of any given tweet, which we operationalize as the likelihood that this tweet provides SA information. Several articles have identified features of tweets that suggest it will provide SA. While future work hopes to consider more complex approaches to determining situational awareness (Corvey et al., 2012; Imran et al., 2013; Munro, 2011; Verma et al., 2011) the present work uses a straightforward, unigram approach similar to that of (Abel et al., 2012). To determine the relevance of a tweet, we simply sum the scores for each keyword found in the tweet. Leveraging further observations from the data, we use this relevance score as a mechanism to perform classification. This is discussed further below.

RESULTS

In this section, we consider results for our three research questions. First, we explore the distribution of benefit scores for our unintuitive keywords. We then report on the overlap between samples of tweets collected using intuitive keywords and those collected using unintuitive keywords. Finally, we explore the quality of tweets found using unintuitive keywords which did not have any intuitive keywords, comparing this to previous findings on tweets found using intuitive keywords (Vieweg, 2012).

Assessing benefit scores

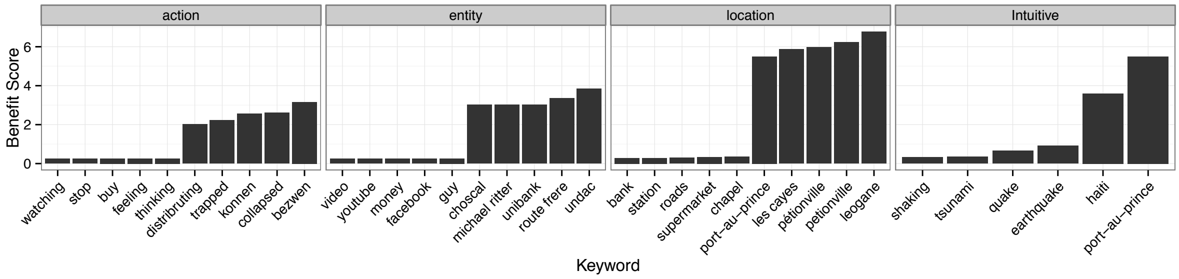

Figure 1 – The top five scoring and bottom five scoring terms for actions, entities and locations. The five terms in the left of each sub-plot are the lowest scoring terms; the five on the right of each sub-plot are the highest scoring terms. We also show scores for any intuitive keywords that also appeared in the Ushahidi reports in the subplot furthest to the right

Figure 1 shows the top five and bottom five scoring terms for the locations, actions and entities found in the Ushahidi reports and the benefit scores for each intuitive keyword that itself appeared in the Ushahidi data. We see that top locations scored higher than top entities, which in turn fair slightly better than actions. We also see that more specific terms tended to be ranked higher than general terms For example, we see that for actions, where almost by definition all terms can be used outside the context of the earthquake, those more specific to the disaster (*bezwen* is Creole for "need", *konnen* for "know") score highly while those centering on less relevant verbs like "feeling" score low.

Figure 1 also shows that the scoring metric provides a conceptually appealing ordering to the intuitive search terms used in previous work that were also in the Ushahidi dataset. Interestingly, "Haiti", while used frequently in the Ushahidi reports, did not score as highly on our metric as "Port-au-Prince" because it was used relatively often in tweets prior to the earthquake. The metric suggests that Haiti was used in more general, global conversation of the earthquake than the term Port-au-Prince. This claim is supported by prior work which suggests people involved in disasters often use localized terminology (Vieweg et al., 2010). Note that any discussion below of tweets found using unintuitive keywords first removes these six terms.

In sum, Figure 1 gives us confidence that our metric is able to identify the different kinds of terms we discussed above. Additionally, results suggest that specific locations in Haiti were likely to have the highest benefit when searching for tweets providing situational awareness, a claim we further in the following subsections. Before continuing, one other point of note is that "Petionville" and "Pétionville", though referring to the same location, both had high benefit scores. While it would have been trivial to disambiguate these two and combine them into one term, one would then only be searching Twitter for the single term used in the disambiguation. This presents a subtle but useful benefit of drawing keywords from the crowd-sourced Ushahidi data. As opposed to official reports or intuition, which promote correctly spelled terms, pulling from crowd-sourced data also allows us to automatically uncover common misspellings that might exist in tweets providing SA.

Finding the overlap between samples

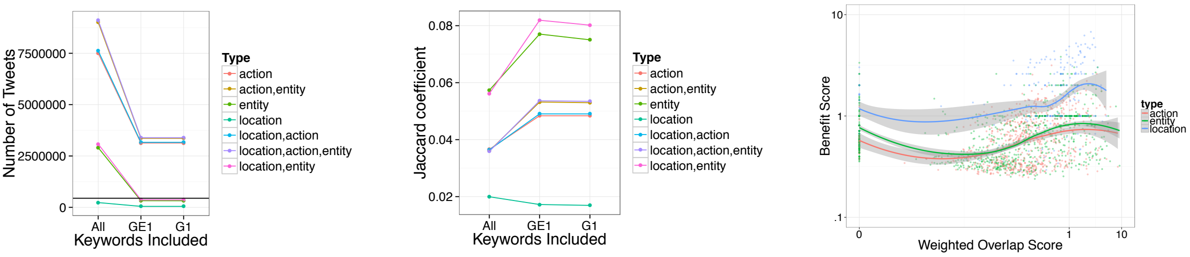

Figure 2 A) The number of tweets (y-axis) collected by searching different types of keywords (color) having different benefit scores (x-axis). The black line near 0 is the size of the set collected when searching for the intuitive keywords. **B)** The same plot as A, except the y-axis has been changed to the jaccard coefficient. **C)** Comparing the benefit score (y-axis) to the weighted overlap score (x-axis, described below) for all unintuitive keywords. Each keyword type is represented by a different color. Each keyword type also has a non-linear regression line computed using a LOESS fit, where grey represents the 95% confidence interval of the regression.

Figure 2A shows the number of tweets in samples collected by using different subsets of the unintuitive keywords. Each colored line represents a different subset of keyword types used in the search, while the three

Process described in (Vieweg, 2012)	**Mirrored steps taken in the present work**
Began with ~4M tweets sent from 1/11-2/2	Began with ~9M tweets sent from 1/12-1/19
Filtered out all tweets where any of a set of 11 noise words were found (e.g. "pray")- see Table 3.1 in (Vieweg, 2012) for the terms used. Now had ~1.4M tweets	Filtered out tweets with the same terms. Now had ~7.5M tweets
Removed any duplicate tweets	Did not remove duplicates, as they made up a small sample of our dataset
Ran the situational awareness classifier from (Verma et al., 2011) on a random sample of ~17% (300K) of the 1.4M tweets due to time complexity of classifier. Resulted in ~230K (75% of the 300K) automatically classified as containing situational awareness	Subsampling was unnecessary, as classification only involves summing term scores. Used classifier described below determine situational awareness
Randomly selected 1,000 tweets from the ~230K	Randomly selected 250 tweets from the ~8K classified as providing SA
Annotators annotated each tweet in these 1,000 tweet subsets as containing information that contributes to situational awareness or not – see (Vieweg, 2012, pp. 163-165 for coding scheme)	Annotation by three annotators

Table 4 Description of the methodology used to ascertain the quality of the tweets found using unintuitive keywords which did not contain intuitive keywords

different points on the x-axis represent restrictions to using all keywords (*All*), only keywords with a benefit score greater than or equal to one (*GE1)* and only keywords with scores greater than one (*G1*). The values for each are compared to the number of tweets found using the intuitive keywords, shown as the black line near the bottom of the figure. Samples that searched using actions and those that utilized all keywords uncovered significantly higher numbers of tweets than intuitive keywords. The figure further shows thatw, as expected, terms with a score of exactly one (those found in only one Ushahidi report and never found in the pre-disaster tweets) added a very small number of tweets to the overall sample.

Figure 2A shows that in almost all cases, there were enough tweets in the sets captured using unintuitive keywords to hold all tweets found using intuitive keywords. However, these two sets of tweets hardly overlapped. Figure 2B uses the same color-coding and x-axis as Figure 2A but instead plots the *jaccard coefficient*, defined as the size of the intersection between two sets divided by the size of their union. In this case, the two sets are the sample of tweets captured using the specified unintuitive keywords and the sample captured using intuitive keywords. The jaccard coefficient can be thought of as a proxy for the overlap between the two sets controlling for their difference. With a maximum jaccard coefficient below .1, Figure 2B shows that unintuitive keywords capture a significantly different set of tweets than intuitive keywords. This shows that keyword choices can severely affect the sample obtained, as the terms found in the Ushahidi reports provided a very different corpora of tweets than those used in previous work.

One question not answered by Figures 2A and 2B, however, is which unintuitive keywords frequently occurred in tweets that also contained intuitive keywords. Figure 2C plots the benefit score of each term versus its *weighted overlap*, defined as $\log(|I_k| + 1) * \frac{|I_k|}{|Tw_k|}$, where $|I_k|$ is the number of times the keyword was found in tweets that had intuitive keywords and $|Tw_k|$ was the number of times the keyword was found in any tweet. Also shown in Figure 2C is a regression curve fitted using a localized nonlinear regression model. The *weighted overlap* and the related curve show the extent to which terms having different benefit scores were found both frequently in general and frequently in tweets having intuitive keywords.

Figure 2C shows that keywords which were never found in the overlap (those having a weighted overlap score of 0) tended to have slightly higher benefit scores than terms found rarely in the overlap. Ignoring these terms, however, we see a positive correlation between benefit score and weighted overlap. Finally, we also observe that locations were significantly more likely to be in the intersection than either entities or actions. These observations give both pros and cons to approaches which begin with a small set of initial intuitive keywords and expand to a larger, more refined set by pulling new terms from tweets captured via this initial set (Li et al., 2013; Lin et al., 2011). On the one hand, we find that for terms found in one or more tweets holding intuitive keywords, higher weighted overlap scores suggested higher benefit scores. On the other had, there were several terms with high benefit scores that were never found in tweets that also contained intuitive keywords. This suggests that drawing search terms from outside of Twitter in addition to capturing new search terms from Twitter itself may improve the ability of these types of approaches, an approach taken by (Abel et al., 2012).

Number of Situationally aware tweets captured

Our third research question asked what the quality was of tweets that do not contain intuitive keywords but that were captured using one or more unintuitive keywords. While a variety of possible avenues could have been taken to approach this problem, we chose to explore quality by comparing results to those in (Vieweg, 2012), who examined the quality of a set of tweets from the Haitian earthquake collected using intuitive keywords.

Table 4 shows the steps taken by Vieweg (2012) and those we have taken to mirror her work as closely as possible. The biggest difference between her efforts and ours was the way in which tweets were classified as having SA information or not. As noted above, we provide a score for each tweet based on a summation of the scores of the keywords within it. We use three observations from qualitative analysis of our data to extend this scoring mechanism into a classifier. First, we note that tweets which scored highly only because they contained a large number of low-scoring terms (e.g. tweets containing several actions and nothing else) tended not to be relevant. We thus use only keywords with a benefit score greater than one, choosing to simply ignore low-scoring terms. Second, as noted above, tweets with only actions were highly unlikely to be relevant to the disaster, as actions provide little contextual information. We thus ignore tweets which do not contain at least one location or entity. Third, the number of tweets that held situational awareness was generally expected to be a small. Our classification mechanism thus only selects tweets scoring in the upper 5% of all observed tweets as having situational awareness.

After completing the steps in Table 4, Vieweg (2012) found that 23% of the 1,000 tweets studied provided situational awareness, 47% were on topic but did not provide situational awareness, and 29.5% were off topic. Due to a lack of time available to annotators not on the study, we chose to annotate only 250 tweets of the 8,000 tweets our classifier determined provided situational awareness. We utilized three coders, all of which were computational social scientists. Tweets not classified the same by all annotators were scored according to the majority vote. The first author of the article acted as the deciding vote when all annotators disagreed, though this occurred in only a single case. Overall, Fleiss' Kappa was 73.5, which is acceptable for the coarse-grained statistics that are of interest in the present work. In our corpus of 250 tweets, 11% were classified as providing situational awareness, 17% were on topic and 72% were off topic. Thus, there were approximately half as many on topic tweets and half as many tweets providing situational awareness as in the sample explored by Vieweg (2012).

There is no doubt the relation between our percentages and Vieweg's (2012) is somewhat biased by the different mechanisms utilized, in particular the use of different classifiers. However, qualitative study of the data supports our finding that tweets not holding intuitive keywords are in general less likely to provide situational awareness and less likely to be on topic than those that do. While a better comparison might have been to also analyze tweets in our own dataset that had intuitive keywords, we determined that such a comparison would not be as meaningful as one that attempted to directly mirrored prior work.

Post-hoc analysis of the classified data also provided two other important points of note. First, we found that tweets classified as having situational awareness scored significantly higher ($F=58.13$, $p < .0001$) than those that did not. This suggests that decreasing the somewhat arbitrary 5% threshold imposed on our classifier could significantly improve its ability to uncover SA tweets, making it simpler and more computationally feasible than the one provided by Verma et al. (2011) and used by Vieweg (2012). Also, our method would provide, to the best of our knowledge, the first unsupervised approach to the classification task of identifying SA information in tweets. However, future work is necessary to further address whether or not our simple unigram model would be as effective in practice. Second, tweets classified as providing SA by the human coders were significantly more likely to have locations ($X^2 = 45.2484$, df=1, $p < .0001$) than those which were not. This further supports the argument presented above for the usefulness of searching for locations.

CONCLUSION

In the present work, we explored the extent to which "intuitive" keywords used in seven previous studies on the 2010 Haitian earthquake were able to capture the bulk of tweets containing situational awareness. We also considered how to identify new, "unintuitive" keywords that might also be useful to search for and the quality of tweets returned from searches using these keywords. Our primary results are three-fold:

> -We presented a novel metric and a novel data source for obtaining new keywords that can be used to search Twitter, showing that the new data source can provide important terms not found via other approaches
>
> -Our metric and several empirical results suggest that locations specific to affected areas of Haiti were useful in uncovering tweets that might have enhanced situational awareness
>
> -We found evidence that situational awareness is sparser in tweets that did not contain search terms used in previous works. However, a significant number of such tweets still existed.

"RT @***: RT @*** Foodmax located in Petion-Ville is now open. Pass the news along to people who live in PAP & need grocer ..."
"Plz HELP orphange La Maison Des Anges 21 Rur Clairsine Tabarre Glasys Maximillien plz call *** NEED WATER #haitiquake #rescuehaiti"
"RT @***: This is a call for help for *** and the 50 children of the Nursery Nid d'Amour. URGENT NEED SUPPLIES Tabarre ..."
@*** Riviere Froide SW of PAP after Leogane, Les Petites Soeurs de Ste Therese,house collapsed,need aid+food+h2o call ***"
"Was the hospital that collapsed in Petion-Ville l'hopital des Petits-Freres? I am looking desperately for a nurse there...***"

Table 5 Five tweets hand-drawn from the top 100 relevant tweets as suggested by our model. Twitter handle names and person names have been replaced with ***

In support of the first and final points above, Table 5 shows five tweets which contain few terms anyone without specific domain knowledge could reasonably have thought to search for, and none of the keywords used in prior work on Haiti. These keywords all contain locations mentioned in the Ushahidi reports, and so could have been uncovered using the approach provided in the present work.

In addition to the described benefits for Twitter, combining these two crowd-sourced data sources could have benefitted Ushahidi. As Ushahidi reports consistently lacked verification, using them to search Twitter leverages a natural link to Twitter's capabilities to provide corroborative information (Palen et al., 2011). In the future, approaches that use Ushahidi reports and other curated data to automatically sift through the Twitter-sphere will allow volunteer aid workers to focus on more trusted sources of data and allow search terms drawn from these sources to uncover both novel and corroborative information from Twitter. One interesting avenue of future work could thus be to better understand how useful Twitter data might have been in corroborating or debunking different Ushahidi reports.

Additional avenues of future work should likely include a more direct comparison of the results of our work to the findings of Vieweg, as a more extensive comparison with a more refined classification mechanism and more experienced data coders is really necessary to obtain conclusive evidence. Finally, while we present our methodology as an interesting means to both obtain and rank new keywords for search, the method has not been tested in the field. In particular, while we have experimented with real-time sampling using these same Ushahidi reports (REMOVED), in these results we consider only post-hoc sampling which is liable to perform better than any system would in real time. We look forward to uncovering how the mechanisms proposed here, along with new methods to make better real-time inferences, can be combined with existing work to increasingly leverage the power of Twitter and other crowd-sourced datasets for disaster relief.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., Tao, K., 2012. Semantics+ filtering+ search= twitcident. exploring information in social web streams, in: Proceedings of the 23rd ACM Conference on Hypertext and Social Media. pp. 285–294.

Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15, 662–679.

Corvey, W.J., Verma, S., Vieweg, S., Palmer, M., Martin, J.H., 2012. Foundations of a Multilayer Annotation Framework for Twitter Communications During Crisis Events, in: Chair), N.C. (Conference, Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turkey.

Harvard Humanitarian Initiative, 2011. Disaster Relief 2.0: The future of information sharing in humanitarian emergencies. UN Foundation & Vodafone Foundation Technology Partnership.

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

Kumar, S., Barbier, G., Abbasi, M.A., Liu, H., 2011. Tweettracker: An analysis tool for humanitarian and disaster relief, in: Fifth International AAAI Conference on Weblogs and Social Media, ICWSM.

Landwehr, P.M., Carley, K.M., 2014. Social Media in Disaster Relief, in: Chu, W.W. (Ed.), Data Mining and Knowledge Discovery for Big Data, Studies in Big Data. Springer Berlin Heidelberg, pp. 225–257.

Li, R., Wang, S., Chen-Chuan, K., 2013. Towards Social Data Platform: Automatic Topic-focused Monitor for Twitter Stream. Proc. VLDB Endow. 6.

Lin, J., Snow, R., Morgan, W., 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11. ACM, New York, NY, USA, pp. 422–429.

MacEachren, A.M., Jaiswal, A., Robinson, A.C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., Blanford, J., 2011. SensePlace2: GeoTwitter analytics support for situational awareness, in: 2011 IEEE Conference on Visual Analytics Science and Technology (VAST). Presented at the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 181 –190.

Morrow, N., Mock, N., Papendieck, A., Kocmich, N., 2011. Independent Evaluation of the Ushahidi Haiti Project (Harvard Humanitarian Initiative Report). Active Learning Network for Accountability and Performance in Humanitarian Action.

Munro, R., 2011. Subword and spatiotemporal models for identifying actionable information in Haitian Kreyol, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 68–77.

Munro, R., 2013. Crowdsourcing and the crisis-affected community. Inf. Retr. 16, 210–266.

Munro, R., Manning, C.D., 2012. Short message communications: users, topics, and in-language processing, in: Proceedings of the 2nd ACM Symposium on Computing for Development, ACM DEV '12. ACM, New York, NY, USA, pp. 4:1–4:10.

O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A., 2010. From tweets to polls: Linking text sentiment to public opinion time series, in: Proceedings of the International AAAI Conference on Weblogs and Social Media. pp. 122–129.

Oh, O., Kwan, K.H., Rao, H.R., 2010. An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter During the Haiti Earthquake 2010. Presented at the International Conference on Information Systems, Saint Louis, Missouri, pp. 1–14.

Palen, L., Vieweg, S., Anderson, K.M., 2011. Supporting "Everyday Analysts" in Safety-and Time-Critical Situations. Inf. Soc. 27, 52–62.

Potts, L., Seitzinger, J., Jones, D., Harrison, A., 2011. Tweeting disaster: hashtag constructions and collisions, in: Proceedings of the 29th ACM International Conference on Design of Communication, SIGDOC '11. ACM, New York, NY, USA, pp. 235–240.

Sarcevic, A., Palen, L., White, J., Starbird, K., Bagdouri, M., Anderson, K., 2012. "Beacons of hope" in decentralized coordination: learning from on-the-ground medical twitterers during the 2010 Haiti earthquake, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, CSCW '12. ACM, New York, NY, USA, pp. 47–56.

Sarter, N.B., Woods, D.D., 1991. Situation awareness: A critical but ill-defined phenomenon. Int. J. Aviat. Psychol. 1, 45–57.

Starbird, K., Munzy, G., Palen, L., 2012. Learning from the Crowd: Collaborative Filtering Techniques for Identifying On-the-Ground Twitters during Mass Disruptions, in: Proceedings of the Conference on Information Systems for Crisis Response and Management (ISCRAM 2012). Vancouver, BC, Canada.

Starbird, K., Palen, L., Hughes, A.L., Vieweg, S., 2010. Chatter on the red: what hazards threat reveals about the social life of microblogged information, in: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, CSCW '10. ACM, New York, NY, USA, pp. 241–250.

Tsagkias, M., de Rijke, M., Weerkamp, W., 2011. Linking online news and social media, in: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11. ACM, New York, NY, USA, pp. 565–574.

Verma, S., Vieweg, S., Corvey, W.J., Palen, L., Martin, J.H., Palmer, M., Schram, A., Anderson, K.M., 2011. Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. Proc ICWSM.

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

Vieweg, S., Hughes, A.L., Starbird, K., Palen, L., 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI '10. ACM, New York, NY, USA, pp. 1079–1088.