An exploration into user-defined keyword sampling from Twitter during disasters

Kenneth Joseph | Kathleen M. Carley

kjoseph | kathleen.carley @cs.cmu.edu
The “search” problem

• Less than .5% of Tweets sampled in a recent paper were found to be “actionable” during a disaster
  – A needle in a haystack (Munro, 2011)

• To make matters worse:
  – You don’t get all the Tweets you’re interested in
    • The needle is in the haystack, but you only get to search through a few bales
  – What you’re interested in is constantly changing
    • The needle has legs

• Human volunteers really shouldn’t be spending their time combing through Twitter

• If you’re a computer, you have to be smart about how you search
The “post-search” problem

• Many have taken the approach of pulling tweets with heavily-used search terms (e.g. Haiti) and then analyzing these as a representative sample to study the uses of Twitter during a disaster
  – E.g. Oh et al, 2010; Hughes and Palen, 2009; Medoza et al, 2010; Starbird and Palen, 2011; Munro, 2010

• A question which has not been asked is, is this an appropriate sampling methodology?

• Just because there is a lot of data, does that mean we can sample, more or less, any way we want?
The research questions

- **RQ1**: can we find Tweets that are “useful” for crisis workers in “real-time”?

- **RQ2**: how different are the tweets we find from those containing characteristic search terms

**The data**
- Gardenhose (about 15% of all Tweets at the time) from the time of through three weeks after the Haitian earthquake of 2010 - ~100M tweets
- We simulate real-time by only allowing ourselves to look at data from previous time periods
Experiment

Our Algorithm

Things to search Twitter for

Get Set of Tweets Back

Random

Things to search Twitter for

Get Set of Tweets Back

Characteristic Search-term based

Things to search Twitter for

Get Set of Tweets Back

Compare
The Algorithm – General Idea

- Develop a method which looks for Tweets that might be “useful” to crisis workers without any work on their part

- How? Find search terms that are relevant and recent using other, more reliable data
  - Recent: our method uses reports from Ushahidi workers to create a dynamic set of search terms to account for concept drift and the differing stages of disasters
  - Relevant: It uses simple unsupervised (boosting) techniques to determine search terms which return relevant results
Ushahidi

- Ushahidi – a crowdsourced platform for mapping situational awareness during a disaster
- Ushahidi got the data from:
  - Mission 4636
  - Mainstream media outlets
  - Twitter
- What did Ushahidi do?
  - Get coordinates, plot incidents on map, and at some points, communicate with Coast Guard
What can we use?

- Categories
  - Ushahidi-volunteers set them, based on a general set of categories for disasters from Red Cross

- Location Data – e.g. “Port-au-Prince”

- The full text description

- The title
  - A more refined version of the text
## Algorithm

### Updating Search Terms

<table>
<thead>
<tr>
<th>Language Model</th>
<th>Search Term Set, t=0</th>
<th>New Incoming Ushahidi Report</th>
<th>Search Term Set, t=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>{}</td>
<td>“Need Water”</td>
<td>{need, water}</td>
</tr>
<tr>
<td>Category</td>
<td>{}</td>
<td>“Water shortage”</td>
<td>{water, shortage}</td>
</tr>
<tr>
<td>Location</td>
<td>{}</td>
<td>“PaP”</td>
<td>{PaP}</td>
</tr>
<tr>
<td>Content</td>
<td>{}</td>
<td>“we are in need of water, please help”</td>
<td>{need, water, help}</td>
</tr>
</tbody>
</table>
The Algorithm
Search Term contextualization

\[
\text{contextualization}(q_i) = \frac{\text{relevance}(q_i) + \text{temporalWeight}(q_i)}{2}
\]

\[
\text{relevance}(q_i) = \frac{n(q_i, C_+)}{n(q_i, C_+) + \sqrt{n(q_i, C_-)}}
\]

\[
\text{temporalWeight}(q_i) = \frac{\text{lastTimeSeenInAReport}}{\text{currentTime}}
\]
The Algorithm

Tweet Score

There is a massive Water Shortage around 5th Street in Port-au-Prince.

The Hotel Montana in Port-au-Prince Has just collapsed!

Oh boy, will there ever be a water shortage in my tomato garden this year!

\{water, shortage\}

\[
p(q_i|t) = \frac{n(q_i, t)}{|t|} (1 - \gamma) + \gamma \frac{n(q_i, T)}{|T|}
\]

\[
p(q_i|t) = \frac{n(q_i, t)}{|t|} (1 - \gamma) + \gamma \frac{n(q_i, T)}{|T|}
\]

\[
\text{score}(t) = \prod_{q_i \in Q \cap t} \sqrt{\text{contextualization}(q_i) \cdot p(t|q_i)}
\]
Algorithm
Combining Models

Oh boy, will there ever be a water shortage in my tomato garden this year!

{need, water}  {water, shortage}  {PaP}  {need, water, help}

No!  Yes!  No!  No!

Not relevant, so update search terms accordingly!

\[
\text{relevance}(q_i) = \frac{n(q_i, C_+)}{n(q_i, C_+) + \sqrt{n(q_i, C_-)}}
\]
Sampling Algorithm - Random

- Randomly sample ~same amount of tweets
- 0.0013% of entire stream
- Allows us to see what our model might be doing well on just because even the dumbest possible model does well on it (it happened...)
Sampling Algorithm

Characteristic Search Terms

<table>
<thead>
<tr>
<th>Work</th>
<th>Search Terms</th>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Munro (2011); Munro and Manning (2012)</td>
<td>#haiti</td>
<td>40,000</td>
</tr>
<tr>
<td>Oh et al. (2010)</td>
<td>#haitiearthquake</td>
<td>962</td>
</tr>
<tr>
<td>Verma et al. (2011); Vieweg (2012)</td>
<td>haiti, earthquake, quake, shaking, 4M tsunami, ouest, Port-au-Prince, tremblement, tremblement de terre (but had to contain “haiti”)</td>
<td></td>
</tr>
<tr>
<td>Sarcevic et al. (2012)</td>
<td>earthquake, Port-au-Prince, Ouest, 3.28M tsunami, haiti, and tremblement</td>
<td></td>
</tr>
</tbody>
</table>

Proj. EPIC Keyword set
Results
Percent of all Tweets selected
Results
Percent of Our Model’s Tweets Captured

![Graph showing the results of percent of tweets captured over time. The graph compares the percentage of tweets captured by our model against three different samples: Ushahidi-Based vs. HaitiEarthquake, Ushahidi-Based vs. Haiti, and Ushahidi-Based vs. Proj. EPIC Keywords. The percentage is shown on the y-axis, and the days from January 14th to January 31st are on the x-axis.](image-url)
### Results

Matching to Proj. EPIC

<table>
<thead>
<tr>
<th>Words</th>
<th>Field</th>
<th>Relevant</th>
<th>Not Relevant</th>
<th>% Reports</th>
<th>Cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td>haiti</td>
<td>Location</td>
<td>152275</td>
<td>302068</td>
<td>0.22</td>
<td>1</td>
</tr>
<tr>
<td>haiti</td>
<td>Title</td>
<td>152207</td>
<td>300428</td>
<td>0.02</td>
<td>0.98</td>
</tr>
<tr>
<td>haiti</td>
<td>Content</td>
<td>152207</td>
<td>300428</td>
<td>0.08</td>
<td>0.95</td>
</tr>
<tr>
<td>earthquake</td>
<td>Content</td>
<td>1318</td>
<td>62151</td>
<td>0.03</td>
<td>0.98</td>
</tr>
<tr>
<td>quake</td>
<td>Content</td>
<td>756</td>
<td>18934</td>
<td>0.01</td>
<td>0.95</td>
</tr>
<tr>
<td>port-au-prince</td>
<td>Location</td>
<td>165</td>
<td>5293</td>
<td>0.19</td>
<td>0.88</td>
</tr>
<tr>
<td>port-au-prince</td>
<td>Content</td>
<td>165</td>
<td>5293</td>
<td>0.07</td>
<td>0.84</td>
</tr>
<tr>
<td>port-au-prince</td>
<td>Title</td>
<td>165</td>
<td>5293</td>
<td>0.02</td>
<td>0.81</td>
</tr>
<tr>
<td>earthquake</td>
<td>Title</td>
<td>1318</td>
<td>62151</td>
<td>&lt;3</td>
<td>0.80</td>
</tr>
<tr>
<td>tremblement</td>
<td>Content</td>
<td>13</td>
<td>225</td>
<td>&lt;3</td>
<td>0.68</td>
</tr>
<tr>
<td>ouest</td>
<td>Location</td>
<td>0</td>
<td>104</td>
<td>&lt;3</td>
<td>0.46</td>
</tr>
<tr>
<td>ouest</td>
<td>Title</td>
<td>0</td>
<td>5</td>
<td>&lt;3</td>
<td>0.46</td>
</tr>
<tr>
<td>quake</td>
<td>Title</td>
<td>0</td>
<td>16553</td>
<td>&lt;3</td>
<td>0.45</td>
</tr>
<tr>
<td>ouest</td>
<td>Content</td>
<td>0</td>
<td>69</td>
<td>&lt;3</td>
<td>0.35</td>
</tr>
<tr>
<td>earthquake</td>
<td>Category</td>
<td>0</td>
<td>19816</td>
<td>&lt;3</td>
<td>0.28</td>
</tr>
<tr>
<td>shaking</td>
<td>Content</td>
<td>5</td>
<td>5617</td>
<td>&lt;3</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Results
The good, the ???, and the ugly
Conclusion

- Much still to be done

- Solid evidence that the model is performing well, however

- Allows Twitter to be searched in real-time, which no direct human interaction

- Validates previous work
Thanks!

- Questions?
Future Work
Results
Hashtag Comparison