On the importance of keywords for the application of Twitter posts for traffic incident detection

Camille Kamga
Anil Yazici
Sandeep Mudigonda
Wei Hao
Nathalie Martinez
Traffic Incidents

- Roadway incidents → 57.9% of the total delay on road networks.
- Improve roadway geometric design for safer driving
- Mitigate incident impacts:
  - 1 min less incident duration → 4-6 min/vehicle delay saving & 9 gal fuel, 0.7 kg HC, 9 kg CO, 1.3 kg NO)
  - Reduce detection and clearance times
    - Gather and disseminate the incident information fastest way possible efficient response
      Crowdsourced social media (Twitter) data can help
    - Harvest the information content of crowd-sourced online Twitter feeds
    - Use as an incident management (IM) support tool

Oak Ridge National Lab Report by Shih-Miao et al., 2004  Texas Transportation Institute, 2012
Use of Social Media

• Web 2.0 → user generated content → everybody is a “reporter”

Social media feeds as information source
• Brand adoption; Political public opinion; “meet up”;
• Monitor disease outbreaks; Disaster information
• Transportation
  • Surveys: policy, demand, etc.
  • Transit service disruptions real-time interaction
  • Potential for extracting real-time information
Transportation Agency Adoptions of Social Media

Iowa DOT

Florida DOT

Utah DOT
Information Extraction from Social Media

• “needle in a haystack” problem (Grant-Muller et al., 2014).
• Natural language form → 80% unstructured (Liu et al., 2011),
  • Ungrammatical, abbreviated
• Approach:
  1. Information retrieval: query-based
  2. Information extraction: text → relevant information
     • “Dictionary” → List of common words → best “candidate” tweets
     • Context dependent, different set for different purposes
     • Lack/ambiguity of context → challenge! (Pereira et al., 2014)
  3. Prediction: extracted information → predict future transportation states
• Most "prominent" (organizational) accounts use incident info from 511, DOT

• Early detection of incidents is possible, for at least few incidents

• Usually from tweets from people (personal accounts)

• Important to distinguish between organizational and personal tweets

⇒ Dictionary!
Proposed Methodology

Twitter Universe

Key words

Twitter API

Initial Crawled Dataset

• Waking up early to beat BQE traffic sucks #offtowork...
• Accident in #Queens...
• Omg a car crashed into ...
• Genius is talent set on fire by courage. - Henry Van ...

Cleaning

tf-idf

Dictionaries weighted words

Potential Dataset ranked for importance

1. Accident in #Queens...
2. Omg a car crashed into ...
3. Waking up early to beat BQE traffic sucks #offtowork...
4. ...
5. ...
10. Genius is talent set on fire by courage. - Henry Van
11. ...

Manually classify raw data into:
Relevant (incident-related) & irrelevant
Organizational account vs. personal accounts

Score tweets using tf-idf “weights” \( \leftarrow \) importance of words

\[
\begin{align*}
    tf(t, d) &= \frac{f(t, d)}{\max \{ f(w, d) : w \in d \}} \\
    idf(t, D) &= \log \left( \frac{N}{\{d \in D : t \in d\}} \right)
\end{align*}
\]
Proposed Methodology

Potential Dataset

1. Accident in #Queens...
2. Omg a car crashed into ...
3. Waking up early to beat BQE traffic sucks #offtowork...
4. ...
5. ...

Manually coded tweets (train)

NB Classifier

Classified Dataset

Classified Geo-coded Dataset

Manually coded tweets (train)

• Naïve-Bayesian (NB) Classifier

\[ P_{NB}(c | d) := \frac{\left( p(c) \sum_{i=1}^{m} p(f | c)^{n(c,d)} \right)}{P(d)} \]

What is the probability that a tweet is relevant given that it includes “car” and “crash”?

• NB for each account type (Organizational vs. personal)
### Geocoding

- < 3% tweets have accurate geo-location

<table>
<thead>
<tr>
<th>Account</th>
<th>Tweet text</th>
<th>Geocode Reported</th>
</tr>
</thead>
<tbody>
<tr>
<td>@TotalTrafficNYC</td>
<td>Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of 6 mins #traffic</td>
<td>-73.9626, -73.9626, -73.6998, -73.6998, 40.5417, ...,</td>
</tr>
<tr>
<td>@sfgiantsfan1</td>
<td>@KTVU there was a high speed crash on Thornton ave in Newark car flipped several times before bursting into flames</td>
<td>-122.0731, -122.0731, -121.9876, -121.9876, 37...</td>
</tr>
<tr>
<td>@511NY</td>
<td>Accident with property damage on #US9 NB at Montrose station rd</td>
<td>-73.9535, -73.9535, -73.9166, -73.9166, 41.2298, ...,</td>
</tr>
</tbody>
</table>
Geocoding

- Regular expressions (ave, pkwy, hwy, st, rd, at, near, between...)
- Hashtags (#Queens)
- Location

<table>
<thead>
<tr>
<th>Tweet text</th>
<th>Geocode Reported</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>@TotalTrafficNYC Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of 6 mins #traffic</td>
<td>-73.9626, -73.9626, -73.6998, -73.6998, 40.5417, ...</td>
<td>Queens, NY</td>
</tr>
<tr>
<td>@KTVU there was a high speed crash on Thornton ave in Newark car flipped several times before bursting into flames</td>
<td>-122.0731, -122.0731, -121.9876, -121.9876, 37...</td>
<td>Newark, CA</td>
</tr>
</tbody>
</table>
Impact of dictionaries

<table>
<thead>
<tr>
<th>Organization accounts</th>
<th>Personal accounts</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;exit&quot; &quot;ave&quot;</td>
<td>&quot;accident&quot;</td>
</tr>
<tr>
<td>&quot;lane&quot;</td>
<td>&quot;block&quot;</td>
</tr>
<tr>
<td>&quot;min&quot;</td>
<td>&quot;delay&quot;</td>
</tr>
<tr>
<td>&quot;right&quot;</td>
<td>&quot;pkwy&quot;</td>
</tr>
<tr>
<td>&quot;crash&quot;</td>
<td>&quot;traffic&quot;</td>
</tr>
<tr>
<td>&quot;close&quot;</td>
<td>&quot;stop&quot;</td>
</tr>
<tr>
<td>&quot;road&quot;</td>
<td>&quot;vehicle&quot;</td>
</tr>
<tr>
<td>&quot;min&quot;</td>
<td>&quot;pkwy&quot;</td>
</tr>
<tr>
<td>&quot;right&quot;</td>
<td>&quot;back&quot;</td>
</tr>
<tr>
<td>&quot;crash&quot;</td>
<td>&quot;clear&quot;</td>
</tr>
<tr>
<td>&quot;close&quot;</td>
<td>&quot;left&quot;</td>
</tr>
<tr>
<td>&quot;road&quot;</td>
<td>&quot;disable&quot;</td>
</tr>
<tr>
<td>&quot;exit&quot; &quot;ave&quot;</td>
<td>&quot;accident&quot;</td>
</tr>
<tr>
<td>&quot;lane&quot;</td>
<td>&quot;block&quot;</td>
</tr>
<tr>
<td>&quot;min&quot;</td>
<td>&quot;delay&quot;</td>
</tr>
<tr>
<td>&quot;right&quot;</td>
<td>&quot;pkwy&quot;</td>
</tr>
<tr>
<td>&quot;crash&quot;</td>
<td>&quot;traffic&quot;</td>
</tr>
<tr>
<td>&quot;close&quot;</td>
<td>&quot;stop&quot;</td>
</tr>
<tr>
<td>&quot;road&quot;</td>
<td>&quot;vehicle&quot;</td>
</tr>
</tbody>
</table>

6900 randomly selected public tweets collected using Twitter API.
Manually coded raw data: incident-related & irrelevant
Organizational vs. personal

\[
\text{Normalized } tfidf(S) = \frac{\sum_{t \text{ for all } d} tfidf(t, d)}{\sum_{t \in S} t}
\]
# Impact of dictionaries

<table>
<thead>
<tr>
<th>Relevant tweet</th>
<th>Account type</th>
<th>Using organizational + personal keywords</th>
<th>Using only organizational keywords</th>
<th>Using only personal keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>State troopers just blocked the ramps leading from route 138 in Canton onto 93 due to serious crash #WCVB</td>
<td>Agency</td>
<td>0.27</td>
<td>0.27</td>
<td>0.8</td>
</tr>
<tr>
<td>Omg a car crashed into the paramus Wendy's @amandabootsy <a href="http://t.co/C4DwTEIyHN">http://t.co/C4DwTEIyHN</a></td>
<td>Personal</td>
<td>0.2</td>
<td>0.16</td>
<td>0.4</td>
</tr>
<tr>
<td>@crosattto it was a bad wreck that a car went straight into the wall and went up in flames. <a href="http://t.co/XCvA7QkAF8">http://t.co/XCvA7QkAF8</a></td>
<td>Personal</td>
<td>0.04</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>car on fire on Lower level of Verrazano Bridge. 🚗🔥🚒 @ Verrazano Bridge Tolls <a href="https://t.co/lpEPEGGXWn">https://t.co/lpEPEGGXWn</a></td>
<td>Personal</td>
<td>0.34</td>
<td>0</td>
<td>1.5</td>
</tr>
</tbody>
</table>
Classification using different dictionaries

- Raw data → 80% training, 20% test
- $\text{NB}_{\text{org}}$ using only organizational dictionary.
- $\text{NB}_{\text{all}}$ using organizational and personal dictionary.
- $\text{NB}_{\text{pers}}$ using only personal dictionary.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy in predicting relevant tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{NB}_{\text{org}}$</td>
<td>75.6%</td>
</tr>
<tr>
<td>$\text{NB}_{\text{all}}$</td>
<td>85.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy in predicting relevant personal tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{NB}_{\text{org}}$</td>
<td>50.5%</td>
</tr>
<tr>
<td>$\text{NB}_{\text{all}}$</td>
<td>54%</td>
</tr>
<tr>
<td>$\text{NB}_{\text{pers}}$</td>
<td>74.4%</td>
</tr>
<tr>
<td>Account</td>
<td>Tweet text</td>
</tr>
<tr>
<td>---------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>@TotalTrafficNYC</td>
<td>Accident cleared in #Queens on The L.I.E. WB at Douglaston Pkwy, stop and go traffic back to x34, delay of 6 mins #traffic</td>
</tr>
</tbody>
</table>
Summary

• All incident information is useful for early detection
• Dictionaries derived from prominent accounts give lesser importance to personal accounts
• Personal dictionaries are more effective in
  • Filtering potentially useful tweets
  • Classification of relevant tweets
• Geocoding requires analysis of regular expressions, hashtags, location of account, neighborhood information
Remarks

• More raw data for personal tweets
• Extra effort for identifying personal & organization (automated)

• IM → incidence, location and time
  • Geo-coding : 3% on all tweets
  • Further text analysis
  • Time of tweet not always incident time
Future Potential

Debris, dead animal $\rightarrow$ Accident prevention!

Information Driven

Operational
Thank you! @nyserda @nysdot #Questions?