Mining Twitter Data to Augment NASA GPM Validation

Bill Teng¹,², Arif Albayrak¹,², George Huffman³, Bruce Vollmer², Carlee Loeser¹,², Jim Acker¹,²
¹NASA GSFC 610.2; ²ADNET Systems, Inc.; ³NASA GSFC 612

NASA CSESP (NNH16ZDA001N)

AGU 2017 Fall Meeting, Dec. 14, 2017
Motivation

Rain event in Los Angeles area

NASA Giovanni visualization of GPM map (left) and time series (right) of the LA area rain event
Organic network of rain gauges

Space-time-varying set of “precipitation tweets”

- Reading the “gauge measurements”
- Develop infrastructure for processing and analyzing tweets
- Enhancing quality of tweets; engaging with “active” participants
- Applying processed tweets to satellite data validation
- Managing tweet data
Tweet processing infrastructure

Filter; pre-process

- Request Query
  - Rain, snow etc.

- Tweet Stream
  (Passive & Hashtaged)

- Ground Truth Data

- Local Tweet Databases

- Training Data Set

Pre-process

Online Process

- Clean Normalize
- Find Precip Type, Intensity
- GPM/TRMM & Tweet

Compare

Co-locate

- GPM/TRMM Data

Manage

- Meta Data

Output

Near Real Time Visualization

Visualize

- States, Counties Statistics Distribution etc.

- Tweet processing infrastructure

Offline Process

- Prepare Training Data Set
- Pre-process
- Train Patterns (Coeff)
- Test

- Learning
  (Tweet type, intensity)
- coefficients
- Training Parameters & data set

Classify

Visualize

- Feed
- Test
- Train

Ground Truth Data

Training Data Set

- Prepare Training Data Set

- Pre-process

- Train Patterns (Coeff)

- Learning
  (Tweet type, intensity)

- coefficients

- Training Parameters & data set
Example tweets

Relevant tweet

Not relevant tweet

Weather station tweet
Anatomy of a tweet (status object)

- **Tweet text**
  - The tweet's unique ID. These IDs are roughly sorted & developers store tweet IDs in a hash table.
  - Text of the tweet. Consecutive duplicate tweets are merged. 140 character max.
  - The date the tweet was created.

- **User info**
  - The author's user name.
  - The author's screen name.
  - The author's location.
  - The author's URL.
  - The author's verified status.
  - The author's time zone.

- **Geo-tag info**
  - The place ID.
  - The type of this place - can be a "neighborhood" or "city."
Approach to crowd-sourcing

- *Not* require participants to explicitly “sign up” to contribute.
- To effectively crowd-source, a large source of crowd is needed.
- Twitter is such a source.
# March 2017 winter storm

## Number of tweets collected


<table>
<thead>
<tr>
<th></th>
<th>Total # tweets</th>
<th># tweets w/ geo-location</th>
<th># tweets w/ geotag (place)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>1,227,390</td>
<td>22,880</td>
<td>34,535</td>
</tr>
<tr>
<td>U.S.</td>
<td>13,269</td>
<td>13,269</td>
<td>20,349</td>
</tr>
</tbody>
</table>
March 2017 snow event
Distribution of tweets
March 2017 snow event
Distribution of tweets
March 2017 winter storm
Observed snowfall & MRMS*-tweet map

*Multi-Radar/Multi-Sensor System (NOAA NSSL)
Comparing tweets with MRMS

- Analysis ongoing; thus far, majority of MRMS-tweet pairs agree (i.e., m1/t1, m0/t0).
- Many “passive” tweets are “regular” (e.g., from amateur weather stations), e.g.,
  - Wind 0.0 km/h N. Barometer 1006.7 mb, Falling. Temperature 8.3 °C. Rain today 8.8mm. Humidity 93%
- Implication for quality of “passive” tweets.
March 2017 snow event
Time-varying # of tweets

~Midnight – 6 am, U.S. east coast
March 2017 snow event
Manhattan
March 2017 snow event

Manhattan

~Midnight – 6 am
# March 2017 snow event

## Frequency of words in tweets

<table>
<thead>
<tr>
<th></th>
<th>Snow</th>
<th>Weather</th>
<th>Storm</th>
<th>Rain</th>
<th>Wind</th>
<th>Stars</th>
<th>Wind Blows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>512,414</td>
<td>229,836</td>
<td>198,798</td>
<td>123,773</td>
<td>113,915</td>
<td>48,789</td>
<td></td>
</tr>
<tr>
<td>U.S. point location only</td>
<td>14,094</td>
<td>4,362</td>
<td>1,850</td>
<td>4,644</td>
<td>9,778</td>
<td>1,686</td>
<td>250</td>
</tr>
</tbody>
</table>

![Term Frequencies Graph](image-url)

- Snow is the most frequently used word in tweets related to the snow event.
August 2017 Hurricane Harvey
Tweets in GeoCollaborate

StormCenter’s GeoCollaborate
August 2017 Hurricane Harvey Tweets in GeoCollaborate
Summary

- Infrastructure is generic, i.e., not specific to a given measurement, social medium, or satellite mission.
- Twitter data have potential for earth science applications.
Questions?

I am tweeting in the rain .. tweeting in the rain ...

.. 'n we're happy again!
Engagement w/ Active Participants

Filter and extract “rain” tweet (Oct. 7, 2016) about Hurricane Matthew.

Active approach to citizen science

“Rain” tweet

Geolocation

Generate rain map from GPM, using NASA Giovanni.

“Reply” to “rain” tweet by tweeting to @NASADISC (special Twitter account created for the experiment) w/ link to rain map.
Managing tweet data
August 2017 Hurricane Harvey
Number of tweets collected

Start: 2017-08-29 00:33:21 – End: 2017-08-29 09:36:02

<table>
<thead>
<tr>
<th></th>
<th>Total # tweets</th>
<th># tweets w/ geo-location</th>
<th># tweets w/ geotag (place)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>1,571,234</td>
<td>6,497</td>
<td>84,103</td>
</tr>
<tr>
<td>U.S.</td>
<td>4,010</td>
<td>78,906</td>
<td></td>
</tr>
</tbody>
</table>

Number of tweets collected
August 2017 Hurricane Harvey
Distribution of tweets