Thinking Critically about Digital Data Collection

Rebekah Tromble
Andreas Storz
Daniela Stockman
What are we observing and why?
Twitter Basics – Means of Collecting the Data

• Application Programming Interfaces (APIs)
  • Firehose – real-time, 100%, cost-prohibitive
  • Streaming – real-time, sample
  • Search/REST – historical, but with significant limitations
  • Archive (not an API) – All tweets since June 2006... sort of...
Understanding the APIs

• Streaming (Keyword queries)
  • Real time capture
  • Can capture up to 1% of global volume – rate limits
    • Issue/event is popular
    • Americans go to sleep/on vacation
• Queries text, hashtags, @mentions, URLs
  • Same as firehose PowerTrack
Understanding the APIs

• Search/Rest (Keyword queries)
  • Historical capture by keyword or username
  • Queries only text
  • Significant limitations:
    • Up to 18,000 tweets over the last ~7-10-day period, whichever limit is reached first.
    • Up to 180 calls every 15 minutes.
    • Captures significantly less than 100% (“top” tweets).
Understanding the APIs

- Archive – not an API
- Not truly a record of all tweets.
- Terms of service require everyone to remove deleted tweets
Previous Research

• Driscoll & Walker, 2014; Morstatter et al, 2013
  • Firehose to Stream
  • Driscoll & Walker:
    • High- and medium-volume events
    • Tweet count comparison only
  • Morstatter et al:
    • Medium-volume event
    • Top hashtags, topic analysis, network centrality measures
Data Collection

- Common set of 6 hashtags, 1-15 October 2014
- Archive (GNIP outage) – several days later, Sifter
  - 556,412 tweets
- Streaming API – TCAT (no rate limits)
  - 470,510 tweets (84.6%)
- Search API – TCAT
  - 255,080 tweets (45.8%)
• Question: Is bias likely to be introduced when data is collected using either API?

• Evaluation:
  • Kendall’s Tau correlation of top mentions and usernames
  • Archive as baseline

• Answer: It’s highly likely.
<table>
<thead>
<tr>
<th>Top #</th>
<th>Archive - Stream</th>
<th>Archive - Search</th>
<th>Archive - Stream</th>
<th>Archive - Search</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.7778</td>
<td>0.2444</td>
<td>0.7333</td>
<td>0.6000</td>
</tr>
<tr>
<td>25</td>
<td>0.8467</td>
<td>0.4667</td>
<td>0.86</td>
<td>0.5776</td>
</tr>
<tr>
<td>50</td>
<td>0.8237</td>
<td>0.6131</td>
<td>0.882</td>
<td>0.6032</td>
</tr>
<tr>
<td>100</td>
<td>0.8179</td>
<td>0.5823</td>
<td>0.9008</td>
<td>0.5702</td>
</tr>
<tr>
<td>250</td>
<td>0.8152</td>
<td>0.5557</td>
<td>0.8528</td>
<td>0.5262</td>
</tr>
<tr>
<td>500</td>
<td>0.8119</td>
<td>0.5145</td>
<td>0.8577</td>
<td>0.5282</td>
</tr>
<tr>
<td>1000</td>
<td>0.8004</td>
<td>0.5249</td>
<td>0.835</td>
<td>0.5376</td>
</tr>
</tbody>
</table>
Question: What factors drive API samples?

Logit regression

- User characteristic variables
  - How prolific? (status count)
  - How popular? (follower count)
  - How engaged? (friend count)

- Tweet characteristic variables
  - Originality? (retweet)
  - Engagement w/ others? (mentions count)
  - Engagement in discourse? (hashtag count)
  - Content richness? (multimedia)
Analysis

- Ran 40 models
  - Step-wise test of interaction effects
  - Simplest proved best.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef</th>
<th>Odds Ratio</th>
<th>Coef</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status count</td>
<td>9.37E-07***</td>
<td>1.0000009</td>
<td>6.05E-07***</td>
<td>1.0000006</td>
</tr>
<tr>
<td>Followers</td>
<td>-9.80E-08***</td>
<td>0.9999999</td>
<td>1.24E-08</td>
<td>1.0000000</td>
</tr>
<tr>
<td>Friends</td>
<td>6.28E-06***</td>
<td>1.0000063</td>
<td>3.77E-07</td>
<td>1.0000004</td>
</tr>
<tr>
<td>Retweet</td>
<td>-3.09E-01***</td>
<td>0.7344833</td>
<td>-5.67E-01***</td>
<td>0.5672836</td>
</tr>
<tr>
<td>Mention count</td>
<td>-4.08E-02***</td>
<td>0.9599965</td>
<td>-3.63E-02***</td>
<td>0.9643693</td>
</tr>
<tr>
<td>Hashtag count</td>
<td>1.24E-01***</td>
<td>1.1323</td>
<td>3.08E-02***</td>
<td>1.0312944</td>
</tr>
<tr>
<td>Multimedia</td>
<td>-5.97E-03</td>
<td>0.9940459</td>
<td>4.58E-01***</td>
<td>1.5816395</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.08E-01***</td>
<td>0.8118603</td>
<td>1.83E+00***</td>
<td>6.2159437</td>
</tr>
<tr>
<td>Variable</td>
<td>Coeff</td>
<td>Odds Ratio</td>
<td>Coeff</td>
<td>Odds Ratio</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
<td>------------</td>
<td>-----------</td>
<td>------------</td>
</tr>
<tr>
<td>Status count</td>
<td>9.37E-07***</td>
<td>1.0000009</td>
<td>6.05E-07***</td>
<td>1.0000006</td>
</tr>
<tr>
<td>Followers</td>
<td>-9.80E-08***</td>
<td>0.9999999</td>
<td>1.24E-08</td>
<td>1.0000000</td>
</tr>
<tr>
<td>Friends</td>
<td>6.28E-06***</td>
<td>1.0000063</td>
<td>3.77E-07</td>
<td>1.0000004</td>
</tr>
<tr>
<td>Retweet</td>
<td>-3.09E-01***</td>
<td>0.7344833</td>
<td>-5.67E-01***</td>
<td>0.5672836</td>
</tr>
<tr>
<td>Mention count</td>
<td>-4.08E-02***</td>
<td>0.9599965</td>
<td>-3.63E-02***</td>
<td>0.9643693</td>
</tr>
<tr>
<td>Hashtag count</td>
<td>1.24E-01***</td>
<td>1.1323</td>
<td>3.08E-02***</td>
<td>1.0312944</td>
</tr>
<tr>
<td>Multimedia</td>
<td>-5.97E-03</td>
<td>0.9940459</td>
<td>4.58E-01***</td>
<td>1.5816395</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.08E-01***</td>
<td>0.8118603</td>
<td>1.83E+00***</td>
<td>6.2159437</td>
</tr>
</tbody>
</table>

**Search**

**Streaming**
<table>
<thead>
<tr>
<th>Variable</th>
<th>Search Coeff</th>
<th>Search Odds Ratio</th>
<th>Streaming Coeff</th>
<th>Streaming Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status count</td>
<td>9.37E-07***</td>
<td>1.0000009</td>
<td>6.05E-07***</td>
<td>1.0000006</td>
</tr>
<tr>
<td>Followers</td>
<td>-9.80E-08***</td>
<td>0.9999999</td>
<td>1.24E-08</td>
<td>1.0000000</td>
</tr>
<tr>
<td>Friends</td>
<td>6.28E-06***</td>
<td>1.0000063</td>
<td>3.77E-07</td>
<td>1.0000004</td>
</tr>
<tr>
<td>Retweet</td>
<td>-3.09E-01***</td>
<td>0.7344833</td>
<td>-5.67E-01***</td>
<td>0.5672836</td>
</tr>
<tr>
<td>Mention count</td>
<td>-4.08E-02***</td>
<td>0.9599965</td>
<td>-3.63E-02***</td>
<td>0.9643693</td>
</tr>
<tr>
<td>Hashtag count</td>
<td>1.24E-01***</td>
<td>1.1323</td>
<td>3.08E-02***</td>
<td>1.0312944</td>
</tr>
<tr>
<td>Multimedia</td>
<td>-5.97E-03</td>
<td>0.9940459</td>
<td>4.58E-01***</td>
<td>1.5816395</td>
</tr>
<tr>
<td>Intercept</td>
<td>-2.08E-01***</td>
<td>0.8118603</td>
<td>1.83E+00***</td>
<td>6.2159437</td>
</tr>
</tbody>
</table>
(Tentative) Conclusions

- Content matters
- User does not
- We are looking at especially “rich” content. This has clear consequences for interpretation.
- Algorithms changing
  - Timeline change likely also affects Search results
  - More emphasis on “verified” user status
New (Even Better) Data

• Trump Inauguration - @realdonaldtrump (very high volume)
• #JointSession (high volume)
• #AHCA (medium volume)
• #jobsreport (low volume)

• PROBLEMS PERSIST ACROSS ALL DATA SETS
Twitter and *Beyond*

- Other APIs even more restrictive
- Don’t know what we don’t know
- The moment of capture matters
One Last Issue...

• When we capture matters.
  • Data decay
  • Metadata
Thank You!

• r.k.tromble@fsw.leidenuniv.nl
• @rebekahktromble