Social media research data

12 February 2016

1-4 pm
The Pleasures and Perils of Studying Social Media

Dr. Katrin Weller
GESIS – Leibniz-Institute for the Social Sciences
Dept. of Computational Social Science
Cologne, Germany

Digital Studies Fellow at John W. Kluge Center
Library of Congress
Washington D.C.

E-Mail: katrin.weller@gesis.org ● Twitter: @kwelle ● Web: www.katrinweller.net
Slides are available at: http://de.slideshare.net/katrinweller
SERIOUSLY? DO THEY NOT REALIZE THAT 99% OF TWEETS ARE WORTHLESS BABBLE THAT READ SOMETHING LIKE ‘JUST WOKE UP. GOING TO STARBUCKS NOW. GETTING LATTE.’

PHOTOS: HTTPS://WWW.FLICKR.COM/SEARCH/?TEXT=COFFEE&LICENSE=4%2C5%2C6%2C9%2C10
Social media research output is growing

No. of publications (Scopus)

(TITLE-ABS-KEY("social media") OR TITLE-ABS-KEY("social web") OR TITLE-ABS-KEY("social software") OR TITLE-ABS-KEY("web 2.0")) AND PUBYEAR > 1999
Pleasures in Studying Social Media
#1: New type of data

- Researchers value social media as a new type of data
- Previously „ephemeral data“ become visible
- Immediate – quick reaction to events
- Structured
- „natural“ data

“What I find really interesting is that structure becomes manifest in internet communication. So it’s the first time in history actually that we can, that social structures between people become manifest within a technology. (...) They become visible, they become crawlable, they become analyzable.”

Social Media Data

- Texts
- Images
- Videos
- Mixed formats / Multimedia
- Connections I (friends, followers)
- Connections II (links/URLs)
- Connections/Actions (likes, favs, comments, downloads)
#2: Various research topics

- User groups
- Events
- Audiences
- Practices
- Information flow
- Influence
- Opinions and sentiments
- Networks
- Interactions
- Predictions

- Language
- Culture
- Political communication
- Activism
- Crisis communication/disaster response
- E-learning
- Health
- Brand communication
#3: Multi-disciplinary environment

- Freedom to explore new approaches
- Multi-method
- Exchange with other disciplines
Scopus: 2000-today by subject area

- Computer Science: 10650, 36%
- Social Sciences: 5542, 19%
- Engineering: 2384
- Medicine: 2288
- Business, Management and Accounting: 2151
- Mathematics: 1535
- Arts and Humanities: 773
- Decision Sciences: 772
- Psychology: 65
- Nursing:
- Economics, Econometrics and Finance:
- Biochemistry, Genetics and Molecular Biology:
- Health Professions:
- Environmental Science:
- Earth and Planetary Sciences:
- Agricultural and Biological Sciences:
- Pharmacology, Toxicology and Pharmaceutics:
- Physics and Astronomy:
- Materials Science:
- Multidisciplinary:
- Neuroscience:
- Immunology and Microbiology:
- Chemical Engineering:
- Veterinary:
- Dentistry:
- Chemistry:
- Energy:
Perils? – Challenges!
#1: Model organisms in social media research?

Social Media Research

Number of publications per year, which mention the respective social media platform’s name in their title. Scopus Title Search. See:
Reasons?

• Accessibility
• Legal framework
• Ethical framework
#2: Comparability?

- Longitudinal studies
- Cross-plattform studies
- Comparisions across countries, populations, topics
- Lack of standards (methods, metadata, benchmark datasets)
Different methods and types of datasets, examples from popular social science papers

<table>
<thead>
<tr>
<th>No.</th>
<th>Method</th>
<th>Domain</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>Analytic: Twitter metrics</td>
<td>Technical</td>
<td>309,740 Twitter users (with followers and tweets)</td>
</tr>
<tr>
<td>[2]</td>
<td>Examination: interviews</td>
<td>Communication</td>
<td>Interviews with 181 Twitter users</td>
</tr>
<tr>
<td>[4]</td>
<td>Analytic: linguistic (sentiment analysis)</td>
<td>Linguistics</td>
<td>20,000 tweets</td>
</tr>
<tr>
<td>[5]</td>
<td>Analytic: linguistic (event detection)</td>
<td>Linguistics</td>
<td>163,500,000 tweets</td>
</tr>
<tr>
<td>[6]</td>
<td>Analytic: linguistic (part of speech)</td>
<td>Linguistics</td>
<td>1,827 annotated tweets</td>
</tr>
<tr>
<td>[7]</td>
<td>Analytic: linguistic (sentiment analysis)</td>
<td>Linguistics</td>
<td>475,000,000 tweets</td>
</tr>
<tr>
<td>[9]</td>
<td>Analytic: linguistic (sentiment analysis)</td>
<td>Linguistics</td>
<td>200,000 annotated tweets</td>
</tr>
<tr>
<td>[10]</td>
<td>Analytic: linguistic (conversation structures)</td>
<td>Linguistics</td>
<td>1.3 million Twitter conversations, with each conversation containing between 2 and 243 posts</td>
</tr>
<tr>
<td>[12]</td>
<td>Analytic: network analysis</td>
<td>Geography</td>
<td>481,248 tweets, 1,953 user pairs</td>
</tr>
<tr>
<td>[13]</td>
<td>Analytic: content analysis, Twitter metrics</td>
<td>Communication</td>
<td>102,500 tweets</td>
</tr>
<tr>
<td>[14]</td>
<td>Examination: experiment</td>
<td>Business</td>
<td>Experiment with 1,677 participants</td>
</tr>
<tr>
<td>[15]</td>
<td>Design and Development: linguistic (method development)</td>
<td>Linguistics</td>
<td>449 tweets sampled from 1.5 GB of Twitter data</td>
</tr>
<tr>
<td>[16]</td>
<td>Examination: survey</td>
<td>Classification</td>
<td>Survey with 505 young American adults</td>
</tr>
<tr>
<td>[18]</td>
<td>Analytic: Twitter metrics, linguistic (sentiment analysis)</td>
<td>Politics</td>
<td>104,003 tweets</td>
</tr>
<tr>
<td>[19]</td>
<td>Analytic: content analysis</td>
<td>Business</td>
<td>93 user profiles and 930 tweets</td>
</tr>
<tr>
<td>[20]</td>
<td>Analytic: content analysis, Twitter metrics</td>
<td>Education</td>
<td>4,574 tweets</td>
</tr>
<tr>
<td>[21]</td>
<td>Examination: survey</td>
<td>Education</td>
<td>Qualitative survey with 11 participants</td>
</tr>
<tr>
<td>[22]</td>
<td>Analytic: network analysis, Twitter metrics</td>
<td>Geography</td>
<td>99,832 tweets</td>
</tr>
<tr>
<td>[23]</td>
<td>Analytic: Twitter metrics, linguistic</td>
<td>Geography</td>
<td>1,535,929,521 tweets from 71,273,997 users</td>
</tr>
<tr>
<td>[25]</td>
<td>Examination: experiment</td>
<td>Education</td>
<td>Two experiments with 125 and 135 students</td>
</tr>
</tbody>
</table>

Table 2. Analysis of methods, domains and datasets in the selected publications.

Example 2008-2013 papers on Twitter and elections: data sources

<table>
<thead>
<tr>
<th>Data source</th>
<th>number</th>
</tr>
</thead>
<tbody>
<tr>
<td>No information</td>
<td>11</td>
</tr>
<tr>
<td>Collected manually from Twitter website (Copy-Paste / Screenshot)</td>
<td>6</td>
</tr>
<tr>
<td>Twitter API (no further information)</td>
<td>8</td>
</tr>
<tr>
<td>Twitter Search API</td>
<td>3</td>
</tr>
<tr>
<td>Twitter Streaming API</td>
<td>1</td>
</tr>
<tr>
<td>Twitter Rest API</td>
<td>1</td>
</tr>
<tr>
<td>Twitter API user timeline</td>
<td>1</td>
</tr>
<tr>
<td>Own program for accessing Twitter APIs</td>
<td>4</td>
</tr>
<tr>
<td>Twitter Gardenhose</td>
<td>1</td>
</tr>
<tr>
<td>Official Reseller (Gnip, DataSift)</td>
<td>3</td>
</tr>
<tr>
<td>YourTwapperKeeper</td>
<td>3</td>
</tr>
<tr>
<td>Other tools (e.g. Topsy)</td>
<td>6</td>
</tr>
<tr>
<td>Received from colleagues</td>
<td>1</td>
</tr>
</tbody>
</table>

"But you can’t make your data available for others to look at, which means both your study can’t really be replicated and it can’t be tested for review. But also it just means your data can’t be made available for other people to say, Ah you have done this with it, I’ll see what I can do with it, (...) There is no open data."

Available datasets

• From individual researchers/groups (sometimes „black market“).
• From conferences: e.g. ICWSM
• Archival institutions: e.g. GESIS (doi:10.4232/1.12319)
Unavailable datasets

Library of Congress' Twitter archive is a huge #FAIL

More than five years on, the library's Twitter archive project is in limbo — with no end in sight.

By NANCY SCOLA | 7/27/15 5:09 PM EDT

The Archive Is Closed

June 3, 2015

By Scott McLemee

Five years ago, this column looked into scholarly potential of the Twitter archive the Library of Congress had recently acquired. That potential was by no means self-evident. The incensed "my tax dollars are being used for this?" comments practically wrote themselves, even without the help of Twitter bots.

For what -- after all -- is the value of a dead tweet? Why would anyone study 140-character messages, for the most part concerning mundane and hyperephemeral topics, with many of them written as if to document the lowest possible levels of functional literacy?

As I wrote at the time, papers by those actually doing the research treated Twitter as one more form of human communication and interaction. The focus was not on the content of any specific message, but on the patterns that emerged when they were analyzed in the


https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research
#5 Moving target

- Platforms and users change
- „Forgotten Features“
  - Favs on Twitter
  - Hashtags on Facebook
  - ...

Lost context: interfaces, look and feel

https://www.flickr.com/photos/jackdorsey/182613360
Lost context: interfaces, look and feel

Facebook on Dec. 25, 2005. Via Internet Archive’s Wayback Machine
Lost context: stories and meaning

#jan25
#6 Data loss

- Deleted tweets
<table>
<thead>
<tr>
<th>ID</th>
<th>Username</th>
<th>Local Time Stamp</th>
<th>Text</th>
<th>Language</th>
<th>Profile</th>
<th>Image</th>
<th>Source</th>
<th>Location</th>
<th>Time Zone</th>
<th>Geo</th>
<th>Hashtags</th>
<th>URLs</th>
</tr>
</thead>
</table>
Format supported by Twitter Terms of services
#7: Ethics and Privacy
Pleasures or Perils?
Questions and Feedback

katrin.weller@gesis.org
@kwelle

http://katrinweller.net