Social Media as Research Data

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
Chances in Social Media Research

- 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 research 2000-today

No. of publications (Scopus)
What is being studied?

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

- Political communication
- Activism
- Crisis communication/disaster response
- E-learning
- Health
- Brand communication
- ...
A new discipline?
Scopus: 2000-today by subject area

- Computer Science: 772
- Social Sciences: 773
- Engineering: 1535
- Medicine: 2151
- Business, Management and Accounting: 2288
- Mathematics: 2384
- Arts and Humanities: 10650
- Decision Sciences: 36%
- Psychology: 5542
- Nursing: 772
- Economics, Econometrics and Finance: 772
- Biochemistry, Genetics and Molecular Biology: 772
- Health Professions: 772
- Environmental Science: 772
- Earth and Planetary Sciences: 772
- Agricultural and Biological Sciences: 772
- Pharmacology, Toxicology and Pharmaceutics: 772
- Physics and Astronomy: 772
- Materials Science: 772
- Multidisciplinary: 772
- Neuroscience: 772
- Immunology and Microbiology: 772
- Chemical Engineering: 772
- Veterinary: 772
- Dentistry: 772
- Chemistry: 772
- Energy: 772
Social Media Data

- Texts
- Images
- Videos
- Mixed formats
- Connections I (friends, followers)
- Connections II (links/URLs)
- Connections/Actions (likes, favs, comments, downloads)

→ Different methods!
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>3</td>
<td>Examination: experiment</td>
<td>Education</td>
<td>Experiment with 123 students.</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>8</td>
<td>Analytic: quantitative (network analysis)</td>
<td>Security</td>
<td>17,803 tweets from 8,616 users + 1st degree network (3,048,360 directed edges, 631,416 unique followers, and 715,198 unique friends)</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>11</td>
<td>Analytic: network analysis, Twitter metrics, clustering, content analysis</td>
<td>Classification</td>
<td>One person’s Twitter network (652 followers, 114 followings). 3,112 tweets.</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 305 young American adults</td>
</tr>
<tr>
<td>17</td>
<td>Design and Development: event detection (method development)</td>
<td>Geography</td>
<td>21,623,947 geo-tagged tweets</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>Analytic: content analysis</td>
<td>Communication</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 91,273,997 users</td>
</tr>
<tr>
<td>24</td>
<td>Analytic: content analysis</td>
<td>Politics</td>
<td>4,869,264 tweets (and 43,378 YouTube URLs)</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.

Social Media Research

Number of publications per year, which mention the respective social media platform’s name in their title. Scopus Title Search. For details: http://kwelle.wordpress.com/2014/04/07/bibliometric-analysis-of-social-media-research/
One of the Challenges: Data Collection and Sharing

“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.”

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>

Archiving Twitter Datasets?
Current approaches
<table>
<thead>
<tr>
<th>ID</th>
<th>User Name</th>
<th>Universal Time</th>
<th>Local Time Stamp</th>
<th>Text</th>
<th>Lang</th>
<th>Profile Image</th>
<th>Source</th>
<th>Location</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>33367655046373376</td>
<td>dimasfest</td>
<td>5/13/2013 3:31</td>
<td>5/13/2013 4:31:08 PM</td>
<td>Finally I've arrived to #WW2013 and querying for the en</td>
<td><a href="http://twitter.com">http://twitter.com</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33367161629071317</td>
<td>hinozok0529</td>
<td>5/13/2013 3:29</td>
<td>5/13/2013 12:29:00 AM</td>
<td>WWW2013の夏の部屋抜けなくて</td>
<td>ja</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>333662718245711787</td>
<td>soyzusistema</td>
<td>5/13/2013 3:11</td>
<td>5/13/2013 12:11:31 AM</td>
<td>WWW2013 Participants, if you're in Online Learning, com</td>
<td>en</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Format supported by Twitter Terms of services
Available datasets

• From individual researchers/groups (sometimes „black market“).
• From conferences: e.g. ICWSM
• Archival institutions: e.g. GESIS (doi:10.4232/1.12319)
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 | 7/15/15 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
Challenges in Archiving Twitter Data
Sources for Challenges

(1) the Twitter Terms of Services
(2) ethical challenges
(3) lack of standard metadata and collection methods
(4) the ever changing nature of Twitter – and Twitter users
Sources for Challenges

(1) the Twitter Terms of Services
(2) ethical challenges
(3) lack of standard metadata and collection methods
(4) the ever changing nature of Twitter – and Twitter users
The changing nature of Twitter in 5 examples
#1

Deleted content
Lost context: interfaces, look and feel
Lost context: stories, meanings
Lost context: user names
#5

URLs and images
Questions and Feedback

katrin.weller@gesis.org
@kwelle

http://katrinweller.net

www.gesis.org/css-wintersymposium
Supplement: some useful references

Tools / Methods for collecting tweets:


[There are much more tools, though. See, e.g. collection at: https://docs.google.com/document/d/1UaERzROI986HqcwrBDLaqGG8X_IYwctj6ek6ryqDOiQ/edit (curated by D. Freelon).]
Supplement: some useful references

Challenges in collecting tweets / data quality:


Supplement: some useful references

Bibliometric studies of Twitter researchers:


Supplement: some useful references

Critical perspectives on data access and inequalities:

Supplement: some useful references

Legal and ethical challenges:

Supplement: some useful references

Twitter’s activities:


• Twitter (no date). Guidelines for using Tweets in broadcast, retrieved January 31, 2015, from https://support.twitter.com/articles/114233.
Supplement: some useful references

Library of Congress’ activities:


Supplement: some useful references

Examples of Twitter datasets shared publicly:

- CrisisLex on Github: [https://github.com/sajao/CrisisLex/tree/master/data/CrisisLexT26/](https://github.com/sajao/CrisisLex/tree/master/data/CrisisLexT26/)
- Hadgu & Jäschke 2014 dataset on Github: [https://github.com/L3S/twitter-researcher](https://github.com/L3S/twitter-researcher)
- Kaczmarek, Lars; Mayr, Philipp (2015): German Bundestag Elections 2013: Twitter usage by electoral candidates. GESIS Data Archive, Cologne. ZA5973 Data file Version 1.0.0, [doi:10.4232/1.12319](https://doi.org/10.4232/1.12319)