Documenting Contemporary Society by Preserving Relevant Information from Twitter

Thomas Risse
L3S Research Center/Leibniz Universität Hannover

Institut für Volkskunde/Kulturanthropologie
Hamburg, 6.1.2016
Social Media

- Important change in the communication on the internet
- Everybody is able to contribute
- Easy to create, share, or exchange information
- Easy to connect with family, friends, colleagues, interesting people
- Can be used everywhere
  - Independent of the location
  - Independent of the medium: Web, Smartphone, Smartwatch, …
- Easy to give feedback
- Different Social Media Applications
  - Social Networks: Facebook, Myspace, LinkedIn, …
  - Blogs: Blogger, Wordpress, …
  - Microblogs: Twitter, Tumblr, identi.ca, Yammer, …
Twitter

- Founded 2006
- Original Idea
  - Sending short message similar to SMS but easier (140 Chars)
  - Receiving messages from interesting people by following them

Sign In Already a member? Welcome back!
Join Twitter Sign up now—it’s fast and free
Activate Already use Twitter on your mobile?

Girl_talk
Twitter is for staying in touch and keeping up with friends no matter where you are or what you’re doing. For some friends you might want instant mobile updates—for others, you can just check the web. Invite your friends to Twitter and decide how connected you want to be.

Twltr-leaf Dom Just got rear ended - small chunk taken out of already imperfect bumper. sigh.

Chris Getting very pissed on free period cocktails. Having fun for two, wishing my love was here to have fun herself

Crystal sanitizing, mouth feels great cos of amazing new floss I bought, and sonicare! gotta love a clean mouth!

rayreadyray Getting coffee at muddys, checking out bikes.

Trapped Ian McKellar Die to family emergency, Wolfmother will not be appearing today
Twitter – An Ambient Friend Following Medium

Until 2006 Twitter asks: “What are you doing?”

Major Content
- Daily Chatters
- Self-promotional
- Spam

@ and # were introduced over time
Twitter - Towards a News Medium for Event-Following

2009 Twitter asks: “What’s happening?”

e.g. a Gossip Machine during a conference

# became important for filtering
Twitter - Towards a News Medium for Event-Following

Citizen Journalism
- Everybody can be a journalist by using Smartphone & Twitter
- E.g. Hudson River Plane Crash 2009

Event Discussions
- 2014 FIFA World Cup semi-final between Brazil and Germany on July 8, 2014
  → 35.6 Million tweets
The Twitter Power

- Its simplicity
- Its speed
- Its domino-effect
Twitter in Numbers

• ~500 million users (320 Million active)
• >340 million tweets per day

What people are talking about

• Pointless babble – 40%
• Conversational – 38%
• Pass-along value – 9%
• Self-promotion – 6%
• Spam – 4%
• News – 4%

Evolutions in different areas

Twitter as a company
- Business model
- Technologies
- End Users
- Industrial use (Analytics)

The Society
- Accepted Communication Medium
- Interaction via Twitter
- Adapts to Twitter (# everywhere)
- Societal needs

The Twitter users
- Originally: ordinary people
- Celebrities, Companies, …
- Use of technologies
  - Tweeting, Re-Tweeting
  - Mobile Usage

The Twitter content
- Different Kinds of Messages
- Content Types (Text, Photo, Video)
- Feedback
  - Reply, Retweets, Likes, …
Twitter itself is evolving
http://twitter.com/public_timeline
User Presentation is changing

Hey there! BarackObama
Twitter is a free service that let's you
phone, or M. Join today to see

Having lunch with milk
families in Indianapolis
for a rally at Indiana
about 5 hours ago from web

In Chapel Hill, Winston-Salem & Hi.
For show times visit: http://rachael)

At a One-Stop Early Vote Rally at T.
More info: http://my.barackoba.com

In New Albany, IN at a town hall meeting
from web

Traveling through PA today & asking
polling location into visit http://pa.

Going to be on The Daily Show with
Hope you watch & reply with your !

Barack Obama
@BarackObama
This account is run by Organizing for
Action staff. Tweets from the President
are signed -bo.

Washington, DC
barackobama.com
Joined March 2007
Born on August 4, 1961

3 followers you know

3 Days Until

0

Follow Barack Obama
Full name
Email
Password

Sign up

Tweets 14.4K
Following 639K
Followers 67.4M
Likes 9
Lists 3

Time to check this one off your bucket list:
ofa.bo/j9W3
The Societal View

• Good representation of our culture and society
• Valuable insights into individuals, groups, and organizations
• Enable an understanding of the public perception of events, people, products, or companies, including the flow of information
• Detailed insights into the day-to-day process of public communication
• Later generations will better understand
  • how topics were spreading
  • how sentiments and opinions were developing
  • better understand the impact of technological developments like Twitter on the evolution of culture and society as it is possible today.

→ Preserving Twitter for future generations is a necessity
Why should we become active?

Short-term usage

• Twitter is an active company with >300 Million customers
• Content remains accessible in the short-term

Mid to Long-term scenario

• Time range: >5 - 10 Years
• IT Industries are moving fast
  Remember: Netscape, Geocities, …
• Is Twitter still existing in 10 years? Bought by another company? Bankrupt?
• What happens with the content?
• Is still easy accessibility provided?
→ We don’t know!!
Digital Preservation

Ensure that digitally encoded information are understandable and usable over the long term

Fundamental Questions

• Preserve what?
• For how long?
• How to test?
• Which people?
• Which organizations?
• …
OAIS Reference Model
Perserve What?

Everything or Subset

Everything
• Huge amounts of data every day (>340 million tweets/day)
• Costs for Twitter access
• Exception: Library of Congress
→ Preserving everything is not practical
→ A subset is more realistic
Selecting a suitable Subset

Depends on
- The aim of the preservation
- Usage scenarios
- Organization and persons involved

Possible solutions
- Sampling of Tweets
  - Broad coverage
  - Already part of Twitter API
- Topical Collection
  - Events (e.g. Olympic Games, Earthquake)
  - Entities (e.g. Angela Merkel, Hamburg)
  - Others (e.g. environment, education)
Which Format to Preserve?

JSON Format

```
{
  "coordinates": null,
  "favorited": false,
  "created_at": "Fri Sep 21 23:40:54 +0000 2012",
  "id_str": "249292149810667520",
  "entities": {
    "urls": [ ]
  },
  "hashtags": [
    {
      "text": "FreeBandNames",
      "indices": [20,34]
    }
  ],
  "in_reply_to_user_id_str": null,
  "text": "Thee Namaste Nerdz. #FreeBandNames",
  "metadata": {
    "iso_language_code": "pl",
    "result_type": "recent"
  },
  "retweet_count": 0,
  "in_reply_to_status_id_str": null,
  "id": 249292149810667520,
  "retweeted": false,
  "user": {
    "profile_sidebar_fill_color": "DDFFCC",
    "profile_sidebar_border_color": "BDDCAD",
    ...
  }
}
```

VS.

HTML Format

```
Why does the train always seem slower when you know you're late?

10:36 AM - 9 Nov 2006

@boelter

© 2015 Twitter About Help Terms Privacy Cookies Ads info
```
Which Format to Preserve?

**JSON Preservation**
- Advantages
  - Easy access via Twitter API
  - Simple Textual Format
  - Open Standard
  - Enables mass processing
- Disadvantage
  - Hardly human readable
  - No visual context of the tweet available

**HTML Preservation**
- HTML Crawl of the Twitter page (→ WebArchive)
- Advantage
  - Preservation of visual representation
- Disadvantage
  - Complex representation with many embedded objects
  - Web crawlers are complex
  - Handling of dynamic pages
  - Complex extraction of content
Is Twitter Content enough?

- A tweet is limited to the most important information
- Can we still understand the meaning and the context in 10 or 20 years?
- We need to make use of all hints we can get to ensure the interpretability
The Web provides more context (2011)

- Gun running from Sudan
- Attack on Copts
- Spam
The Web provides more Context (2016)

RT @RSSEGYPTcom: #Egypt #Jan25
أحكام عيد من 3-10 سنوات ضد النبلاء والشرطة بالإكراه وحظر التجوال
http://dlvr.it/KCHz8 #tahrir
about 1 hour ago via EgyTweets

RT @RSSEGYPTcom: #Egypt #Jan25
http://dlvr.it/KCHyYd #tahrir
about 1 hour ago via EgyTweets

RT @ScoutfunLA: comes rough, tough like an elephant tusk. Ya head rush, fly like Egyptian musk...
about 1 hour ago via EgyTweets

RT @therealmonaldaven: DailyViewEgypt: Egypt sheltered trucks bringing arms from Sudan http://dlvr.it/4x9Q6d
about 1 hour ago via EgyTweets

RT @AMasryAlYoum_E: Armed forces attacked sleeping #Copts, say Coptic leaders. http://dlvr.it/4eQOv4 #Albwen #Egypt
about 1 hour ago via EgyTweets

RT @techsynd: Intel Buys Egypt-Based SyQSD to Re-Boost its 4G LTE Efforts. http://dlvr.it/4brzHm
about 1 hour ago via EgyTweets

Server not found

Pale Moon can't find the server at www.techsynd.com.

- Check the address for typing errors such as www.example.com instead of www.example.com.
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Pale Moon is permitted to access the Web.
Another Example: Speak2Tweet

Speak To Tweet
@speak2tweet
Click the link in each tweet to hear a voice tweet.
 Joined January 2011

Tweets Tweets & replies
Speak To Tweet @speak2tweet · Jul 14
voice-to-tweet: g.co/s2t/pgk9

Speak To Tweet @speak2tweet · Jul 4
voice-to-tweet: g.co/s2t/ctcd

Speak To Tweet @speak2tweet · Jun 29
voice-to-tweet: g.co/s2t/k2bh

Speak To Tweet @speak2tweet · Jun 13
voice-to-tweet: g.co/s2t/rmp7

Speak To Tweet @speak2tweet · May 26
voice-to-tweet: g.co/s2t/7Hz
Requirement for Web / Twitter Crawling

Topical Dimension
- Crawl intention are mainly focused around events and rarely around entities
- What is the intention of the user/researcher?
- Easy monitoring by the researcher and possibility to correct

Flexible Crawling Strategies
- Shallow observation crawls
- Focused crawls with prioritization (e.g. PageRank and/or semantics)

Social Web Crawling
- General interest with different media focus
- Integrated with Web crawler to capture the full context

Authenticity
- See a web page as the user saw the page (e.g. including ads and tweets at that time point)

Context and Provenance
- Demographics of sites
- Documentation of crawl specification and history
Current approach: Collect, then crawl

- Social Media: scalable access only through API
- Requires special client programming
- Not supported by typical crawlers
- 2 step workaround: collect Tweets, extract URLs, crawl URLs
- Huge delay between time of post and time of crawling!
Web changes in response to current events

Internet Archive June 18\textsuperscript{th}, 2015, 3:17 vs. 17:06 (same day)

Freshness of Integrated Crawling

- Rapid change of linked URLs (~few hours)

- Analysis of Social Media collection can change based on crawl time!

→ URLs in Social Media need to be crawled soon after posting
Integrated Crawling approach

- API client and Web crawler cooperate through shared queue
- URLs in Tweets are inserted early in the queue to ensure timely crawling
- Crawl continues also from tweeted URLs
Integrated crawling with the L3S iCrawl System

L3S iCrawl System (under development)
- Learning the intention of the crawl
- Integration of Web and Social Media Crawling
- Content based monitoring of the crawl process
Crisis in Ukraine: In depth news, commentary and analysis ...

Ukraine News + Nachrichten / Ukraine aktuell
http://news.feed-reader.net/386-ukraine.html

Welcome to Ukraine
http://ukraine.org/

The Interpreter auf Twitter: "Ukraine Live: NATO Secretary General Tells Russian Foreign Minister To End Support for Ukrainian Rebels http://t.co/c8EX5TkrP6"
https://twitter.com/Interpreter_Mag/status/560661031866768409

Keywords: Retweets, Medien, Schießen Schießen, Übersetzung, Liste Liste, Land, Folgen Folge, Vodafone, Verifierter Liste
Entities: NATO, Secretary General, Russia, Foreign Minister, Ukraine, Russian, Foreign, Minister, To, End, Support, Ukrainian, Rebels, http://t.co/c8EX5TkrP6

Crisis in Ukraine: Get in depth coverage of the country’s political crisis and Russia’s de facto annexation of Crimea, from FT.com

Ukraine Live: NATO Secretary General Tells Russian Foreign Minister To End Support for Ukrainian Rebels http://t.co/c8EX5TkrP6

RT @MillerMENA: NATO is wasting its breath -- neither Russia nor its proxies want peace in Ukraine http://t.co/MW3VdAVcDU
Example for Integrated Crawling

Twitter #Ukraine Feed

Lars Olberg @MissileMonitor - 5h
Media: #Ukraine may buy French #Exocet anti-ship missiles, missilethreat.com/media-ukraine

Michael Baskin @michael300e - 5h
#Ukraine accuses #Russia of sending in dozens of tanks, weapons into Russian separatists held areas bin.ws/?sm Witt P

Fox News
Ukraine accuses Russia of sending in dozens of tanks, weapons into... Ukraine on Friday accused Russia of sending dozens of tanks and other heavy weapons into its rebel-controlled easternmost regions.

View on web

(High Page Relevance)

(Medium Page Relevance)

UK1 1 http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/ 1.00
UK2 1 http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/ 1.00
UK3 x http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/ 0.40
UK4 γ http://missilethreat.com/turkey-missile-talks-france-china-disagreements-erdogan/ 0.05

…

Crawler Queue

ID Batch URL Priority
UK1 1 http://www.foxnews.com/world/2014/11/07/ukraine-accuses-russia-sending-in-dozens-tanks-other-heavy-weapons-into-rebel/ 1.00
UK2 1 http://missilethreat.com/media-ukraine-may-buy-french-exocet-anti-ship-missiles/ 1.00
UK3 x http://missilethreat.com/us-led-strikes-hit-group-oil-sites-2nd-day/ 0.40
UK4 γ http://missilethreat.com/turkey-missile-talks-france-china-disagreements-erdogan/ 0.05

…

Web Link Extracted URL
The Result

Collection of contextualized Tweets

- The tweets
  - A machine readable representation in JSON
  - Embedded objects (e.g. pictures)
- The context of the tweets
  - The profile of the person
  - The visual representation of the user page
  - Topics defined by hashtags
  - The answering and re-tweeting chain
  - Additional information by following Web links

Next Steps

- Preparation for usage
  - Cataloguing
  - Information extraction and enrichment
  - Big Data Analytics
- Ingest into the preservation system
  - Preparing for long-term preservation
  → Not part of this talk
Cataloguing

Focused/topical crawl campaigns result in large number of collections
- Catalogue ensures visibility on a collection level
- Meta-information
  - Provide insights into the collection
  - Creation time frame
  - Content description
  - Technical information, e.g. format, size, …
- Provenance information
  - Creation process and documentation
  - Selection strategies and terms
  - Used technologies
  - …
Information Extraction and Enrichment

Aim

• Provide efficient access to the vast amount of information in a semantically meaningful way

• Examples for questions to be answered
  • How did people talk about the issue or event?
  • How are opinions distributed in relation to demographic user data?
  • Who are the most active Twitter users?
  • Who are the opinion leaders?
  • Where did they come from?
  • What did they talk about?
  • How has the public opinion on a key person evolved?
Extracting Entities

- Paris Hilton
- Bataclan theatre in Paris
- City or Person
  - Paris, France
  - Paris, Texas
  - Paris Hilton
- Hotel or Family
  - Hilton Hotel Chain
  - Family Hilton
- Paris, France

Paris Hilton Online @OnlineHilton · Dec 12
Paris Hilton visits the Bataclan to pay tribute to the victims of the terror attacks in the French capital.

dailymail.co.uk/tvshowbiz/arti…
Enriching Entities

Paris Hilton

Bataclan theatre in Paris

Paris, France

https://en.wikipedia.org/wiki/Paris_Hilton

https://en.wikipedia.org/wiki/Bataclan_(theatre)

https://en.wikipedia.org/wiki/Paris
Event Extraction


Bataclan, Paris

Event: terror attack

Paris Hilton

Paris, France

Paris Hilton visits the Bataclan to pay tribute to the victims of the terror attacks in the French capital.

Issues With Analysing Social Media

- informal use of language and style (abbreviations, slang, domain-specific terms, and spelling and grammatical errors)
- NLP tools are develop for standard language → Lower quality on social media
  - Affects all phases of analysis and extraction
- Approaches
  - Retraining of NLP components
  - replacing common abbreviations with their full words (e.g., tnx = thanks)
  - adding lists of emoticons
  - recognising artefacts such as smileys, @mentions, and hashtags separately
Extraction Results

Extraction Summary
• Entities: Paris Hilton (Person), Bataclan (Location), Paris (Location)
• Event: Paris Attack (13 November 2015)
• Keywords: tribute, victims
• Sentiments: none (only by analyzing the linked page)

Metadata
• Author: OnlineHilton
• Date: 12. December 2015
• Links: http://www.dailymail.co.uk/tvshowbiz/article-3356119/Paris-Hilton....html
• Etc.
Archive Access

Different user groups with different needs
• Not the typical Web Search Engine user

Explorative Search
• Hugh amounts of data
• E.g. reading 35 Million tweets about the World Cup semi-final between Brazil and Germany is not realistic

Large scale analysis
• Aggregated views on content
• Visual analytics
• Drill-down to the interesting tweets
Simple Filtering – FireMe!

Haters

Sexual Intercourse

http://fireme.l3s.uni-hannover.de/
Sentiment Analysis: „Refugees“

https://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/
Drill down into details

Sentiment: pleasant, confident

Topics: refugees

Heatmap: active, alert, excited, elated

Tag Cloud: refugees

Timeline: 07/01/16

Map: Drill down into details

Affinity: pleasant, high confidence

Tweets: refugee, border, police, abuse

What Do I Do?

Type a keyword to see related topics, sentiment, and tweets determined by your keyword. Move your mouse over a tweet to see its text.

Ref Text:

**Tweet Details**

**ResRussia**

Dec 23, 1:08pm: #Russia Bulgarian Border Police Accused of Abusing Refugees: Interviews with aid workers and dozens of refugees...
[https://t.co/8IIyPMRnH](https://t.co/8IIyPMRnH)

$v = 3.12, a = 5.10$

border, $v = [\mu: 4.9, \sigma: 1.49], a = [\mu: 4.07, \sigma: 1.98], f_a = 50$

accused, $v = [\mu: 2.92, \sigma: 1.55], a = [\mu: 4.79, \sigma: 2.51], f_a = 50$

abuse, $v = [\mu: 1.8, \sigma: 1.23], a = [\mu: 0.63, \sigma: 2.7], f_a = 10$

**Tweet Details**

**Henkdoorvoogel**

Dec 23, 1:09pm: RT @OpenDoorsUK: Are you a fantastic fundraiser? Could you raise £10K for Christian refugees? Check out the #Muskathlon...
[https://t.co/Fejyf...](https://t.co/Fejyf...)

$v = 5.96, a = 6.53$

fantastic, $v = [\mu: 7.78, \sigma: 1.00], a = [\mu: 6.23, \sigma: 2.73], f_a = 50$

raise, $v = [\mu: 6.74, \sigma: 1.21], a = [\mu: 7.17, \sigma: 2.06], f_a = 50$

check, $v = [\mu: 6.1, \sigma: 1.53], a = [\mu: 6.1, \sigma: 2.19], f_a = 50$
Topic and Sentiment Analysis: „Refugees“
Analysing Rumours Spread on Twitter

http://www.theguardian.com/uk/interactive/2011/dec/07/london-riots-twitter
Long Term Interpretability

Named Entity Evolution Recognition
Language changes over time!

- Our language is DYNAMIC and changes with our culture, politics, technology, social media, etc.
- The Twitter limitations force users to use a “creative” language

- Different spellings over time
- New words or acronyms are introduced
- Words change their meanings
- Names of concepts are changing

In long-term digital archives → Problems!

<table>
<thead>
<tr>
<th>St. Petersburg</th>
<th>Petrograd</th>
<th>Leningrad</th>
<th>St. Petersburg</th>
<th>St. Petersburg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1914</td>
<td>1924</td>
<td>1991</td>
<td>today</td>
<td>t</td>
</tr>
</tbody>
</table>
Problem: Finding Documents

A scholar writing a thesis about "Pope Benedikt":

→ "I want to know more about Pope Benedikt"
Named Entity Evolution

- Named Entities (NE): people, places, companies...
- Characteristics of Named Entity Evolution (NEE)
  - Same thing but different terms over time
  - Change occurs over short periods of time
  - Small or no concept shift
  - Announced to the public repeatedly
- Goal: Find method for named entity evolution recognition independent from external knowledge sources
In his latest address to American bishops visiting Rome, Pope Benedict XVI stressed that Catholic educators should remain true to the faith—a reminder issued just in time for another tense season of commencement addresses. No, the pope did not mention Georgetown University by name when discussing the Catholic campus culture wars.

**Evaluation Results**

- Burst detection found total 73% of all change periods
- High recall for unsupervised method
- Machine Learning boosts precision
- Data Set: [http://www.l3s.de/neer-dataset/](http://www.l3s.de/neer-dataset/)
Research Infrastructure on Social Mining & Big Data Ecosystem

- **Big Data Analytics & Social Mining** as a tool to measure, understand and possibly predict human behavior.
- **Research infrastructure (RI) for ethic-sensitive scientific discoveries** and advanced applications of social data mining.
- **Integrating key national infrastructures and centers of excellence**: CNR & Uni Pisa, Uni Hannover, Uni Sheffield, Fraunhofer, ETH Zürich, …
  - Social Data: Social Media, Movement data, Shopping data, …
  - Tools: Natural language processing, text mining, data mining, visual analytics
- **Communities**: analysts and social informatics researchers, Economists, social science and humanities researchers, journalists, policy and law makers, related communities, the public as a whole
- **Runtime**: 2015 - 2019
- **EU H2020 Research Infrastructures project #654024**
Conclusions

Twitter Preservation
• Twitter can provide more than short term views
• Twitter preservation enable long term studies

Twitter Crawling
• Twitter crawls should include the context
  • Context of the content
  • Visual presentation

Twitter Archive Access
• Different from classic archives
• Aggregated view are starting points
• Big Data Analytics
• Language changes

Legal Aspects
• Twitter rules hinder the official collection of tweets
• Exception: National Libraries and Archives
Thank You!

Dr. Thomas Risse
Forschungszentrum L3S
Leibniz Universität Hannover
Appelstrasse 9a
30167 Hannover, Germany

E-Mail: risse@L3S.de
Telefon: +49-511-762 17764
Telefax: +49-511-762 17779