computational social media

lecture 3: tweeting

part 1

daniel gatica-perez

20.03.2020
announcements

reading #2 will be presented today:
M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, PNAS, 2013

assignment #2 will be given today

projects will be discussed today
this lecture

a human-centric view of twitter

1. introduction:
2. twitter users & uses
3. large-scale human behavior & real-world events
1. introduction
**mission**
“to give everyone the power to create and share ideas and information instantly, without barriers”

<table>
<thead>
<tr>
<th>Category</th>
<th>2014</th>
<th>2016</th>
<th>2018 &amp; later</th>
</tr>
</thead>
<tbody>
<tr>
<td>monthly active users</td>
<td>241M</td>
<td>320M</td>
<td>N/A</td>
</tr>
<tr>
<td>tweets sent per day</td>
<td>500M</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>active users on mobile</td>
<td>76%</td>
<td>80%</td>
<td>N/A</td>
</tr>
<tr>
<td>accounts outside the US</td>
<td>77%</td>
<td>79%</td>
<td>N/A</td>
</tr>
<tr>
<td>supported languages</td>
<td>35+</td>
<td>35+</td>
<td>N/A</td>
</tr>
<tr>
<td>employees</td>
<td>2700</td>
<td>3900</td>
<td>N/A</td>
</tr>
</tbody>
</table>
"an echo chamber of random chatter"

“140 characters: between an SMS (with larger audience) and a blog (but less cumbersome)”

280 characters (Nov 2017)

J. van Dijck The culture of connectivity, Oxford University Press, 2013
just setting up my twttr
before twitter...

古池や蛙飛び込む水の音
ふるいけやかわずとびこむみずのおと

old pond . . .
a frog leaps in
water’s sound

Bashō (17th century)

The Dinosaur

On waking, the dinosaur was still there.

Augusto Monterroso (20th century)

http://en.wikipedia.org/wiki/Haiku

http://es.wikipedia.org/wiki/Microrrelato
Send the following telegram, subject to the terms on back hereof, which are hereby agreed to:

Dr. Enrico Fermi  
Institute of Nuclear Studies  
University of Chicago  
Chicago, Illinois

RESERVATIONS MADE AT RITTENHOUSE CLUB, 1811 WALNUT STREET FOR BOTH EVENINGS.
CHAINS REQUIRED
20 MILES AHEAD
EXPECT DELAYS
2. twitter users and uses
what is twitter made of?

follow (2006)
users subscribe to other users’ tweets

hashtag # (2007, official 2009)
words articulating a topic or event
allow for search and clustering

retweet RT (2007, official 2009)
repost tweets towards one’s followers
enables trends by retweeting

https://about.twitter.com/press/brand-assets
https://about.twitter.com/milestones
J. van Dijck The culture of connectivity, Oxford University Press, 2013
hashtags

- link **strangers** into larger conversations
- facilitate **impromptu** interactions
- not directed communication but a **stream**
- enable the **emergence** of trending topics
Chris Messina
@chrismessina

how do you feel about using # (pound) for groups. As in #barcamp [msg]?

Eric Rice
@ericrice

ReTweet: jmalthus @spin Yes! Web2.0 is about social media, and guess what people like to be social about? Themselves. Social Narcissism

https://twitter.com/ericrice
geolocalized tweets in Switzerland
who uses twitter?
users and usage

2006: older professional users in business and news
2009: shift to younger adults, then mainstream

from social network to information network

---

**tool for communication**
- everyday small talk
- (citizen) journalism
- political grassroots activism
- emergencies and disasters
- community participation
- fake news

**tool for self-promotion**
- celebrities, stars
- politicians
- enables organization/management of fans/audiences/voters

---

“the impulse to make life a publicly annotated experience has blurred the distinction between advertising and self-expression, marketing and identity” (Hagan 2011)

---

J. van Dijck The culture of connectivity, Oxford University Press, 2013
CCDF: Complementary Cumulative Distribution Function
CCDF(x) = 1 − F(x) = P(X>x)

“connection with very low expectation”: weak ties (Murthy, 2013)

low reciprocity, highly asymmetric links (Kwak, 2010)

“77.9% of user pairs with any link between them are connected one-way.”

“67.6% of users are not followed by any of their followings. For these users Twitter is rather a source of information than a social networking site”

3. large-scale human behavior & real-world events
twitter and human mood

S. A. Golder and M. W. Macy, Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures, Science, 30 September 2011, Vol. 333 no. 6051 pp. 1878-1881
mood

“a conscious state of mind or predominant emotion” (Merriam-Webster dictionary)

“a temporary state of mind or feeling” (Oxford dictionary)
goal and data

positive affect (PA):
enthusiasm, delight, activeness, alertness

negative affect (NA):
distress, fear, anger, guilt

PA and NA are independent dimensions
low PA: absence of positive feelings, not presence of negative ones

**goal:** study variations in PA & NA over time of day, day of week, and world region using longitudinal twitter data

* 2.4 million twitter users worldwide
* 509 million tweets
* up to 400 public messages per user
* all users had at least 25 messages
* average: 212 tweets/user
* period: 02.2008 and 01.2010
* only english speakers
extraction of positive affect (PA) and negative affect (NA)

Linguistic Inquiry & Word Count (LIWC)
- categories related to psychological constructs and personal concerns
- word count per category

PA

NA
What is LIWC?

Linguistic Inquiry and Word Count (LIWC) is a text analysis software program designed by James W. Pennebaker, Roger J. Booth, and Martha E. Francis. LIWC calculates the degree to which people use different categories of words across a wide array of texts, including emails, speeches, poems, or transcribed daily speech. With a click of a button, you can determine the degree any text uses positive or negative emotions, self-references, causal words, and 70 other language dimensions.

The LIWC program can analyze hundreds of standard ASCII text files or Microsoft Word documents in seconds. The LIWC2007 program also allows you to build your own dictionaries to analyze dimensions of language specifically relevant to your interests. The Macintosh version of LIWC2007 has a feature that will highlight in color all the words found in a particular file when it is analyzed. With the Macintosh version, users can also create dictionaries that include literal phrases (e.g. 'you know') as well as individual words and word stems.

The student version of LIWC, LIWClite7, only analyzes plain text files using the LIWC2007 and earlier LIWC2001 dictionaries. LIWClite7 is the student version that is ideal for people with limited text analysis needs.
LIWC (2)

dictionary of 4,500 words

each word belongs to one or more categories

“agree” is part of: affect, positive emotions and assent

64 word categories (excluding punctuations)

http://www.liwc.net/

http://liwc.wpengine.com/
positive emotion category words
negative emotion category words
measurements

\[
PA_u(h) = \frac{\|PA \text{WORDS}_u(h)\|}{\|WORDS_u(h)\|}
\]

where \( h \in H \) and \( H = \{0 \ldots 167\} \), or the 168 hours of the week (24 hours/day * 7 days). The measure for NA was computed similarly, as were the measures taken over 24 hours.

Between-individual variation captures how individuals differ from one another in their baseline affect regardless of the time of day or day of week. It is simply the individual's mean affect across all hours:

\[
BPA_u = \overline{PA_u} = \frac{1}{\|H\|} \sum_{h \in H} PA_u(h)
\]

The within-individual PA score for a person-hour measures the signed difference between the person's score that hour and their baseline as defined in (2). Within-individual scores are comparable across people because individuals' baseline tendencies toward being upbeat or downbeat have been removed, leaving only the change over time that is within each individual:

\[
WPA_u(h) = PA_u(h) - BPA_u + \frac{1}{\|UH\|} \sum_{(u,h) \in UH} PA_u(h)
\]

where \((u, h)\) pairs indicate user-hours and \(UH\) is the set of all such pairs in the dataset.\(^1\)
Fig. 1 Hourly changes in individual affect broken down by day of the week (top, PA; bottom, NA). Each series shows mean affect (black lines) and 95% confidence interval (colored regions).
Fig. 2 Hourly changes in individual affect in four English-speaking regions. Each series shows mean affect (black lines) and 95% confidence interval (colored regions).
A. Mislove et al., Pulse of the Nation: US mood throughout the day inferred from Twitter (2010)
http://www.ccs.neu.edu/home/amislove/twittermood
twitter practical session
(presented by Trung)
what to remember

**twitter as an information network**
- brevity as a value
- a network of weak links
- low reciprocity, highly asymmetric

**large-scale human behavior**
- traces of human states like mood
- language in twitter is short yet informative
- beware of biased data
questions?
daniel.gatica-perez@epfl.ch