Social Media Metadata as Sociolinguistic Evidence

Jacob Eisenstein
@jacobbeisenstein

Georgia Institute of Technology

January 7, 2016
Social media

NASA has confirmed that the asteroid is headed directly for us.

...Yes, a question?

What role has Social Media played in this asteroid’s orbit?

*Sigh*

Has Twitter changed the way we respond to asteroid threats?

Well, it’s made the press conference questions stupider.

Fascinating!

What about Facebook?
- Not a register, genre, or dialect.
- A diverse array of communicative platforms
  - Mostly text
  - Largely informal
- Several platforms support large-scale data acquisition
Why you should care about social media

- Scale
- Variation and change
- Metadata

Social media often includes sociolinguistically relevant metadata:
- social networks
- demographics
Scale: quantification of rare phenomena

(From Jack Grieve at NWAV44)
Scale: discovery of new variables

lbvs: laughing but very serious

- i wanna rent a hotel room just to swim lbvs
- tell ur momma 2 buy me a car lbvs

(Eisenstein, 2015b)
Why you should care about social media

- Scale
- Variation and change
- **Metadata**
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics
Social networks in sociolinguistics

- Reveal the microstructure of language change (Labov, 2001; Dodsworth, 2014).
- Modulate the influence of demographic categories (Milroy & Milroy, 1992).
- Define local communities of linguistic practice (Eckert, 2000).
Social networks in social media

Social media platforms offer a number of forms of metadata that capture social networks.

Articulated network Explicitly-defined connections; undirected in Facebook, directed in Twitter.

Behavioral network Inferred from conversational interactions, such as replies or mentions. (danah boyd, 2006)
Social networks on Twitter

- Twitter users often follow 1000s of other users.
- Mention networks are smaller, and arguably more socially meaningful.
- Twitter query rate limiting makes mention network much easier to obtain.

(Huberman et al., 2008)
Case study 1: Online audience design

Do social media users modulate the standardness of their language according to the audience?

- Social media services offer digital affordances to control the likely audience for a message.
- The connection between language variation and these affordances can reveal the social meaning of each.
- Geographical variables (like lbvs) that are reserved for more local audiences may be more stable.
Our full programme will follow in couple of days! We’re very excited about it - so many great talks!
Hashtag-initial
Addressed
Logistic regression

- **Dependent variable**: does the tweet contain a non-standard, geographically-specific word (e.g., lbvs, hella, jawn)
- **Predictors**
  - **Message type**: broadcast, addressed, #-initial
  - **Controls**: message length, author statistics
Small audience $\rightarrow$ less standard language
Distinguishing local ties

To distinguish local audiences:

- Use GPS metadata to identify author locations
- Associate metro $m$ with user $u$ if $u$ is @-mentioned by:
  - at least three users within metro $m$;
  - nobody outside metro $m$. 

(Sadilek et al., 2012)
Distinguishing local ties

To distinguish local audiences:

- Use GPS metadata to identify author locations
- Associate metro $m$ with user $u$ if $u$ is @-mentioned by:
  - at least three users within metro $m$;
  - nobody outside metro $m$.

The social network lets us impute the locations of unknown users from the 1-2% of users who reveal their GPS! (Sadilek et al., 2012)
Local audience $\rightarrow$ less standard language

Local ties make non-standard language even more likely.
Local audience → less standard language

More mentions by users in **same** metro area

More mentions by users in **other** metro areas

Messages containing local variable

Messages not containing local variable
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics
# Demographics in social media

<table>
<thead>
<tr>
<th>Category</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>All internet users</td>
<td>18%</td>
<td>23%*</td>
</tr>
<tr>
<td>Men</td>
<td>17</td>
<td>24*</td>
</tr>
<tr>
<td>Women</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>White, Non-Hispanic</td>
<td>16</td>
<td>21*</td>
</tr>
<tr>
<td>Black, Non-Hispanic</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Hispanic</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>18-29</td>
<td>31</td>
<td>37</td>
</tr>
<tr>
<td>30-49</td>
<td>19</td>
<td>25</td>
</tr>
<tr>
<td>50-64</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>65+</td>
<td>5</td>
<td>10*</td>
</tr>
<tr>
<td>High school grad or less</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Some college</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>College+ (n=685)</td>
<td>18</td>
<td>30*</td>
</tr>
<tr>
<td>Less than $30,000/yr</td>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>$30,000-$49,999</td>
<td>18</td>
<td>21</td>
</tr>
<tr>
<td>$50,000-$74,999</td>
<td>15</td>
<td>27*</td>
</tr>
<tr>
<td>$75,000+</td>
<td>19</td>
<td>27*</td>
</tr>
<tr>
<td>Urban</td>
<td>18</td>
<td>25*</td>
</tr>
<tr>
<td>Suburban</td>
<td>19</td>
<td>23</td>
</tr>
<tr>
<td>Rural</td>
<td>11</td>
<td>17</td>
</tr>
</tbody>
</table>
Demographics in social media

Women Are More Likely to Use Pinterest, Facebook and Instagram, While Online Forums Are Popular Among Men

% of online adults by gender who use the following social media and discussion sites

<table>
<thead>
<tr>
<th>Platform</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>66</td>
<td>77</td>
</tr>
<tr>
<td>Pinterest</td>
<td>16</td>
<td>44</td>
</tr>
<tr>
<td>Instagram</td>
<td>24</td>
<td>31</td>
</tr>
<tr>
<td>LinkedIn</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>Twitter</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>reddit, Digg or Slashdot</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Tumblr</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Pew Research Center surveys conducted March 17-April 12, 2015.

PEW RESEARCH CENTER
Demographics from geography

- The U.S. Census collects detailed demographics for multiple levels of geographic detail.

- For each geotagged message, treat the average census demographics as a predictor.

- Possible objections:
  - Census regions are too demographically heterogeneous.
  - People move around too much.
  - Social media users are not a representative sample of their census region.
Case study 2: Dialect in writing

Some non-standard spellings hint at dialectal pronunciations:

(ing)

(-t,-d)

How do the demographic properties of these spellings align with the demographics of the associated pronunciations? (Eisenstein, 2015a)
## G-deletion

<table>
<thead>
<tr>
<th></th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>.227</td>
<td>.200</td>
<td>89,173</td>
</tr>
<tr>
<td>Noun</td>
<td>-.013</td>
<td>.083</td>
<td>18,756</td>
</tr>
<tr>
<td>Adjective</td>
<td>-.213</td>
<td>.149</td>
<td>4,964</td>
</tr>
<tr>
<td>monosyllable</td>
<td>-2.57</td>
<td>.001</td>
<td>108,804</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>.178</strong></td>
<td></td>
<td><strong>112,893</strong></td>
</tr>
</tbody>
</table>

("high"/"low" = top/bottom quartile)
## G-deletion

<table>
<thead>
<tr>
<th></th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>.227</td>
<td>.200</td>
<td>89,173</td>
</tr>
<tr>
<td>Noun</td>
<td>-.013</td>
<td>.083</td>
<td>18,756</td>
</tr>
<tr>
<td>Adjective</td>
<td>-.213</td>
<td>.149</td>
<td>4,964</td>
</tr>
<tr>
<td>monosyllable</td>
<td>-2.57</td>
<td>.001</td>
<td>108,804</td>
</tr>
<tr>
<td>High Euro-Am county</td>
<td>-.194</td>
<td>.117</td>
<td>28,017</td>
</tr>
<tr>
<td>High Afro-Am county</td>
<td>.145</td>
<td>.241</td>
<td>27,022</td>
</tr>
<tr>
<td>High pop density county</td>
<td>.055</td>
<td>.228</td>
<td>27,773</td>
</tr>
<tr>
<td>Low pop density county</td>
<td>-.017</td>
<td>.144</td>
<td>28,228</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>.178</td>
<td></td>
<td>112,893</td>
</tr>
</tbody>
</table>

("high" / "low" = top/bottom quartile)
-t,-d deletion

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel succeeding context</td>
<td>.483</td>
<td>-.066</td>
<td>.385</td>
<td>9,004</td>
</tr>
<tr>
<td>Total</td>
<td>.423</td>
<td></td>
<td></td>
<td>89,174</td>
</tr>
</tbody>
</table>
-t,-d deletion

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel succeeding context</td>
<td>.483</td>
<td>-.066</td>
<td>.385</td>
<td>9,004</td>
</tr>
<tr>
<td>@-message</td>
<td>.519</td>
<td>.075</td>
<td>.436</td>
<td>35,240</td>
</tr>
<tr>
<td>Total</td>
<td>.423</td>
<td></td>
<td></td>
<td>89,174</td>
</tr>
</tbody>
</table>
## -t,-d deletion

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Log odds</th>
<th>%</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel succeeding context</td>
<td>.483</td>
<td>-.066</td>
<td>.385</td>
<td>9,004</td>
</tr>
<tr>
<td>@-message</td>
<td>.519</td>
<td>.075</td>
<td>.436</td>
<td>35,240</td>
</tr>
<tr>
<td>High Euro-Am county</td>
<td>.422</td>
<td>-.313</td>
<td>.311</td>
<td>19,992</td>
</tr>
<tr>
<td>High Afro-Am county</td>
<td>.516</td>
<td>.065</td>
<td>.508</td>
<td>19,854</td>
</tr>
<tr>
<td>High income county</td>
<td>.473</td>
<td>-.107</td>
<td>.388</td>
<td>20,653</td>
</tr>
<tr>
<td>Medium income county</td>
<td>.495</td>
<td>.019</td>
<td>.406</td>
<td>43,135</td>
</tr>
<tr>
<td>Low income county</td>
<td>.532</td>
<td>.127</td>
<td>.482</td>
<td>25,386</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>.423</strong></td>
<td><strong>.89,174</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Dialect in writing

- Both non-standard spellings are...
  - less frequent in counties with many European Americans;
  - more frequent in counties with many African Americans.
- (ing) is more frequent in urban counties.
- (-t,-d) is more frequent in low-income counties.

Next questions:
- Do these observations generalize to the writers themselves?
- Does linguistic systematicity vary with demographics? (Labov, 2007)
Dialect in writing

- Both non-standard spellings are...
  - less frequent in counties with many European Americans;
  - more frequent in counties with many African Americans.
- (ing) is more frequent in urban counties.
- (-t,-d) is more frequent in low-income counties.

Next questions:
- Do these observations generalize to the writers themselves?
- Does linguistic systematicity vary with demographics? (Labov, 2007)
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics
Why you should care about social media

- Scale
- Variation and change
- Metadata
  - Social media often includes sociolinguistically relevant metadata:
    - social networks
    - demographics
Case study 3: gender and social networks

We started with a painfully simple idea: train a classifier to predict author gender from language and social network features (Bamman et al., 2014)

- Prior work shows that text predicts author gender (Koppel et al., 2002)
- Social networks are often assortative with respect to gender (McPherson et al., 2001).
- Can we build a better classifier by putting these two features together?
Data

14,464 Twitter users from 2011 (56% male)

- geolocation in USA
- must use 50 of 1000 most frequent words
- no more than 1000 follow connections

In total: 9.2M tweets, from January to June 2011
The U.S. Social Security Administration collects yearly statistics on given names and gender.

\[ P(\text{age, gender} \mid \text{name}) = \frac{\text{count}(\text{age, gender, name})}{\text{count}(\text{name})} \]
Demographics from names

- Limited to names that occur at least 1000 times in the Census data.
  - \( \sim 9,000 \) names in total
  - Most infrequent: Cherylann, Kailin, Zeno
- The median author’s name is 99.6% homogeneous by gender.
- 95% of all authors have a name that is at least 85% associated with one gender.
Behavioral network induced from **mutual** @-mentions

- Mentions must occur over a period of at least two weeks.
- Moderate gender assortativity:
  - Women have 58% female friends.
  - Men have 67% male friends.
Automatic classification

Logistic regression from bag-of-words features gives 88% accuracy.

- This is similar to prior work (Koppel et al., 2002; Burger et al., 2011).
- Will social network homophily help fix the remaining errors?
Adding social network features

Logistic regression, 10-fold cross-validation:
- Text alone: 88% accuracy

With $\geq 1000$ words per author, adding network info does not improve accuracy. Why not?
Adding social network features

Logistic regression, 10-fold cross-validation:

- Text alone: 88% accuracy
- Text+network: 88% accurate

With ≥ 1000 words per author, adding network info does not improve accuracy. Why not?
Adding social network features

Logistic regression, 10-fold cross-validation:
- Text alone: 88% accuracy
- Text+network: 88% accurate

With \( \geq 1000 \) words per author, adding network info does not improve accuracy. **Why not?**
female authors

percent male social network

classifier confidence male decile
Why social network features don’t help

<table>
<thead>
<tr>
<th>text vs network correlation</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>female authors</td>
<td>0.38 (0.35 ≤ r ≤ 0.40)</td>
</tr>
<tr>
<td>male authors</td>
<td>0.33 (0.30 ≤ r ≤ 0.36)</td>
</tr>
</tbody>
</table>
Why social network features don’t help

<table>
<thead>
<tr>
<th></th>
<th>text vs network correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>female authors</td>
<td>0.38 (0.35 ≤ r ≤ 0.40)</td>
</tr>
<tr>
<td>male authors</td>
<td>0.33 (0.30 ≤ r ≤ 0.36)</td>
</tr>
</tbody>
</table>

- Language and social network are correlated even after controlling for author gender.
- Rather than seeing linguistic features as revealing the author’s gender, they reveal an attitude towards gender.
Summary of case studies

1. Audience design on the Twitter social network
2. Demographic profiles of spelling variables
3. Linking language, gender, and social networks
# Language at the intersections

Top words at the intersection of place, age, and gender:

<table>
<thead>
<tr>
<th>Age</th>
<th>Sex</th>
<th>New York</th>
<th>Dallas</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-17</td>
<td>F</td>
<td>niall, ilysm, hemmings, stalk, ily</td>
<td>fanuary, idol, Imbo, lowkey, jonas</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>ight, technique, kisses, lesbian,</td>
<td>homies, daniels, oomf, teenager, brah</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dicks</td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>F</td>
<td>roses, castle, hmmm, chem, sinking</td>
<td>socially, coma, hubby, bra, swimming</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>drunken, manhattan, spoiler,</td>
<td>harden, watt, astros, rockets, mavs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>guardians, gonna</td>
<td></td>
</tr>
<tr>
<td>30-39</td>
<td>F</td>
<td>suite, nyc, colleagues, york, por-</td>
<td>astros, sophia, recommendations,</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>tugal</td>
<td>houston, prepping</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mets, effectively, cruz, founder,</td>
<td>texans, rockets, embarrassment, tcu,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>knicks</td>
<td>mississippi</td>
</tr>
<tr>
<td>40+</td>
<td>F</td>
<td>cultural, affected, encouraged,</td>
<td>determine, islam, rejoice, psalm,</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>proverb, unhappy</td>
<td>responsibility</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reuters, investors, shares, law-</td>
<td>mph, wazers, houston, tx, harris</td>
</tr>
<tr>
<td></td>
<td></td>
<td>suit, theaters</td>
<td></td>
</tr>
</tbody>
</table>

(Pavalanathan & Eisenstein, 2015)
Methodological pros and cons

Pros:

▶ Orders of magnitude more data
▶ Rapid changes in progress
▶ Metadata, especially social networks
▶ Informal communication with real social stakes

Cons:

▶ Biased, non-representative population sample
▶ Language is embedded in socio-technical systems that makes purely linguistic inferences difficult.
Acknowledgments

Thanks to my collaborators:

- David Bamman (UC Berkeley)
- Brendan O’Connor (UMass)
- Umashanthi Pavalanathan (Georgia Tech)
- Tyler Schnoebelen (Idibon)

Funding support:

- National Science Foundation
- Air Force Office of Scientific Research
Demographics from profile pictures?

One approach: use crowd workers to identify author demographics from profile images (e.g., Cheong & Lee, 2009).

- But this relies on essentializist and reductive models of race and gender (Sharma, 2013)
- High agreement ≠ high accuracy.
Demographics from profile pictures?

One approach: use crowd workers to identify author demographics from profile images (e.g., Cheong & Lee, 2009).


▶ But this relies on essentializist and reductive models of race and gender (Sharma, 2013)

▶ High agreement ≠ high accuracy.
References


