BAD LANGUAGE!
...on the INTERNET!!

Boom! Ya ur website suxx bro

...dats why pluto is pluto it can neva be a star

michelle obama great. job. and. whit all my. respect she. look. great. congrats. to. her.

What can we do about it?
Why don't they just write NORMALLY??
Can our software ever ADAPT???

I now h v an iphone
How does language go bad?

Illiteracy? No. (Tagliamonte and Denis 2008; Drouin and Davis 2009)

Length limits? (probably not)

Hardware input constraints? (Gouws et al 2011)

Social variables

- Non-standard language does *identity work*, signaling authenticity, solidarity, etc.
- Social variation is usually inhibited in written language, but social media is less regulated than other written genres.
Abnormal | Normal

Normalization

imma → I'm gonna → I am going to

dats why pluto is pluto → That's why pluto is pluto

Boom! Ya ur website suxx bro → Boom! Yes, your website sucks, brother.

Irredeemably abnormal? | Whose norm?

hella | flavor
wtf | Ain't
jawn | flavour
Lots of work on “X-for-Twitter” using domain adaptation.

- POS: Gimpel et al 2011
- Parsing: Foster et al 2011

Is social media a domain? Is Twitter?
Coherence over time

- **Goal**: measure the linguistic coherence of Twitter
- **Data**: million-word samples at each month and hour
- **Measure**: relative proportion of OOV bigrams

Social media language is changing continuously.

- We cannot annotate our way out of the Bad Language problem.
- Any annotated dataset rapidly becomes stale.
Coherence across media

Twitter is self-similar, but...
OOV rate increases significantly across usage scenarios

<table>
<thead>
<tr>
<th></th>
<th>Tw-June</th>
<th>Tw-@</th>
<th>Tw-#</th>
<th>Blog-body</th>
<th>Blog-comment</th>
<th>Infinite-Jest</th>
<th>PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tw-June</td>
<td>27.8</td>
<td>28.7</td>
<td>29.3</td>
<td>47.1</td>
<td>48.6</td>
<td>54.0</td>
<td>63.9</td>
</tr>
<tr>
<td>Tw-@</td>
<td>25.9</td>
<td></td>
<td></td>
<td>47.8</td>
<td>49.9</td>
<td>56.3</td>
<td>66.4</td>
</tr>
<tr>
<td>Tw-#</td>
<td>29.8</td>
<td>33.4</td>
<td></td>
<td>49.6</td>
<td>51.0</td>
<td>54.7</td>
<td>66.2</td>
</tr>
<tr>
<td>Blog-body</td>
<td>41.9</td>
<td>44.1</td>
<td>43.8</td>
<td>27.2</td>
<td>49.1</td>
<td></td>
<td>48.0</td>
</tr>
<tr>
<td>Blog-comment</td>
<td>47.4</td>
<td>49.6</td>
<td>49.2</td>
<td>30.2</td>
<td>53.0</td>
<td></td>
<td>48.4</td>
</tr>
<tr>
<td>Infinite-Jest</td>
<td>49.4</td>
<td>51.1</td>
<td>49.9</td>
<td>48.3</td>
<td>47.4</td>
<td></td>
<td>55.5</td>
</tr>
<tr>
<td>PTB</td>
<td>72.2</td>
<td>73.1</td>
<td>72.7</td>
<td>64.5</td>
<td>61.9</td>
<td></td>
<td>71.9</td>
</tr>
</tbody>
</table>

PTB is the clear outlier
most OOV tokens in almost every comparison

- Tw-June: randomly-selected messages from June 2011
- Tw-@: messages beginning with a username mention
- Tw-#: messages beginning with a hashtag
- Blog-body: posts from 2008 political blogs (Yano et al 2009)
- Blog-comment: from 2008 political blogs (Yano et al 2009)
- Infinite-Jest: the 1996 novel (Wallace 2012)
- PTB: section 2-21