From TweetGenie to the Apocalypse

Theo Meder
Dong Nguyen
Rilana Gravel
Corpus collection

- **User selection**
  - Dutch Twitter users were sampled by searching on the query ‘het’. An additional set was collected by randomly sampling from the followers/followees of the initial set of users.
  - Sampled in August 2012.

- **Annotation**
  - Two annotators were employed to annotate properties such as gender, age, life stages, geographic location, type of avatar etc.
  - 3185 users annotated.
## Language use on Twitter: Previous Research

<table>
<thead>
<tr>
<th>Features</th>
<th>Males</th>
<th>Females</th>
<th>‘Younger’ users</th>
<th>‘Older’ users</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Articles, prepositions, hyperlinks, numbers, ...</td>
<td>Personal pronouns, conjunctions, blog words, ...</td>
<td>Blog words, negative emotion words, personal pronouns, ...</td>
<td>Hyperlinks, positive emotion words, numbers, ...</td>
</tr>
<tr>
<td>Topics</td>
<td>Religion, politics, business, internet, technology, ...</td>
<td>Conversation, home, fun, romance, ‘me now’, ...</td>
<td>Conversation, home, fun, romance ‘me now’, ...</td>
<td>Religion, politics, business, internet, technology, ...</td>
</tr>
</tbody>
</table>
Gender- and age-based differences
Gender and age prediction

- A system that predicts the gender, exact age, age categories and life stages using logistic and linear regression (exact age) based on 200 tweets of a user.

- **Automatic** versus **manual** prediction.
  - Gender: Comparable performance
  - Age: Automatic system performed better than humans, while being much faster. For example, humans were on average 4.8 years of the exact age, the automatic system 4.1 years.

- Harder to estimate the age for older persons.
Red = women, Blue = men
How old am I?

- @USER Woensdagochtend 15 augustus start het landelijke CDA met haar regiotour op Goeree-Overflakkee i.s.m. @USER.

- (@USER On Wednesday morning, the 15th of August starts the national CDA with its tour through the region in Goeree-Overflakkee in collaboration with @USER)

- RT @USER: Vanmiddag met @USERMENTION gezellig bij @USERMENTION een wijnje gedaan en naar de Emmauskerk #Middelharnis geweest. Mooie dag zo!

- (RT @USER: Had fun this afternoon had wine at @USER with @USER and went to the Emmauschurch #Middelharnis. Beautiful day!)
How old am I?

- The user is a 19-year old student
- Words related to politics are highly ranked features associated with older users
- Prepositions, conventional punctuation, formal abbreviations, conservative vocabulary and for example talking about wine lead to the classification of the user as a 38-year old employee

- People can choose to emphasize aspects of their identity that are not related to gender or age
- Deviation from the style and contents from a person’s peers: this makes the automatic prediction of gender and age more difficult
TweetGenie

www.tweetgenie.nl

(online soon)
TweetGenie: examples

- TheoMeder (M, 52)
- Antalvdb (M, 43)
- Rensbod (M, 48)
- Estherouwehand (F, 37)
- Claudiadebreij (F, 38)
- Ialalalalinder (Linda Duits, F, 36)
- MarcMarieH (M, 48)
- GeertWildersPVV (M, 49) vs GeertWilders (fake account)
Ik kan aan de hand van je tweets je leeftijd en geslacht voorspellen.
Uitproberen? Vul dan hieronder je tweet account in (je tweets moeten publiekelijk beschikbaar zijn)

Volgens mij is theomeder een... man en circa 39 jaar oud.

Deel je resultaat op Twitter

Had ik het geslacht goed?  
Ja  Nee

Goede leeftijd:  
Verzend
The Twitter Apocalypse

- Tweets in total: 52,848 (retweets: 32,055 = 61%)

Number of tweets about 21-12-2012
The Twitter Apocalypse (2)

- The most popular retweet (4050 x): “Those who believe that the world will end on December 21, please donate all your money to my bank account by December 20.”
The Twitter Apocalypse (3)

- Belief and religion
- LIWC: Linguistic Inquiry and Word Count

- Massive fear or media hype? How many Dutch people on Twitter really feared the world would end?
The Twitter Apocalypse (4)

- 2012-12-16: @USER: Pam and I are afraid that on December 21 the world will end =”("
The Twitter Apocalypse (4)

- 2012-12-16: @USER: Pam and I are afraid that on December 21 the world will end =”(

- Ca. 500 tweets of genuine fear = 1 %
Conclusions

- Linguistic research / TweetGenie: gender and age
- Apocalypse: sentiment, popularity, belief
- Conduits and micro-variation
- Representation of self
References


- Nguyen, Dong, Noah Smith and Carolyn Rosé. 2011. Author Age Prediction from Text using Linear Regression. Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities at ACL.

- Nguyen, Dong, Rilana Gravel, Dolf Trieschnigg & Theo Meder: "'How Old Do You Think I Am?" A Study of Language and Age in Twitter', in: ICWSM 2013 (submitted, accepted).

