Social media data
and its potential for official statistics

Piet J.H. Daas
and Marco Puts, Joep Burger and Martijn Tennekes
Why social media?

– Dutch are very active on social media!
– Around 70% according to a recent study
– Readily available
– Potential source of information on what goes on in 'Dutch' society (active on social media)
– As an auxiliary source? In addition to survey and admin. data

– Any other possibilities?
– Explore!

Map by Eric Fischer (via Fast Company)
1. Compare Twitter topics and theme's CBS publications
2. Social media sentiment and consumer confidence
3. ‘Measure’ other basic emotions in social media
4. Social cohesion and Twitter (for a municipality)
5. Selectivity: background characteristics of Twitter users
6. Event detection on Dutch highways
7. More on the way ...
2) Sentiment indicator

- Determine sentiment in *public* Dutch social media messages
  - Huge amounts of Facebook and Twitter messages
  - \( \frac{\# pos - \# neg}{\# total} \) (day/week/month)
- High correlation (> 0.8) with consumer confidence index
- Both series cointegrate (\( \rightarrow \) strong association)

- How's the situation now?
  - For > 5 years of data (Jan. 2009 - March 2015)
‘Sentiment’ indicator for NL (beta-version)

Based on the average sentiment of public Dutch Facebook and Twitter messages
3) Basic emotions in Social Media

A number of basic emotions

- Happy
- Scared
- Angry
- Tender
- Excited

Examples of associated emotions:

- Happy: Ecstatic, Energetic, Aroused, Bouncy, Nervous, Perky, Ansty
- Scared: Irritated, Resentful, Miffed, Upset, Mad, Furious, Raging
- Angry: Tense, Irresolute, Irritated, Frightened, Riled, Stricken, Terrified
First results
5) Selectivity: Twitter user characteristics

– Only a part of the Dutch are active on Twitter
– If we want to use this source we need more info
– By determining their ‘background’ characteristics
  – Such as gender, age, income, level of education etc.

– What are the possibilities?
  - Feature extraction is the way to go
    • For gender
Piet Daas
@pietdaas

Researcher, Big Data scientist and father of 3.

Eindhoven
about.me/pietdaas
Joined February 2010

4) Picture

3) Messages

40 Photos and videos

TWEETS
1,665
FOLLOWING
74
FOLLOWERS
175
FAVORITES
81
LISTS
1

Tweets
Tweets & replies
Photos & videos

Piet Daas retweeted
Big Data Network @BigDataNetwork · 16h

#BigData News » Apache Spark jumps on the R bandwagon: Apache Spark, the big data processi... bit.ly/1w2lqJo via @BigDataNetwork

Piet Daas retweeted
Dr. Diego Kuonen @DiegoKuonen · Feb 21

Is redoing scientific research the best way to find truth? sciencenews.org/article/redoin...
#Science #BigData #Reproducibility
Studied a Twitter sample

- From a list of Dutch Twitter users (~330,000)
- A random sample of 1000 unique ids was drawn

- Of the sample:
  - 844 profiles still existed
    - 844 had a name
    - 583 provided a short bio
    - 473 created ‘tweets’
    - 804 had a ‘non-default’ picture
  - 409 Men (49%)
  - 282 Women (33%)
  - 153 ‘Others’ (18%)
    - companies, organizations, dogs, cats, ‘bots’..
Gender findings: 1) First name

Used Dutch ‘Voornamenbank’ website (First name database)
Score between 0 and 1 (female – male); 676 of 844 (80%) names were registered
Unknown names scored -1 (usually companies/organizations)
Gender findings: 2) Short bio

- If a short bio is provided
  - Quite a number of people mention their ‘position’ in the family
    • Mother, father, papa, mama, ‘son of’, etc.
  - Sometimes also occupations are mentioned that reflect the gender (‘studente’)
  - 155 of 583 (27%) indicated their gender in short bio
  - Need to check both English and Dutch texts
Gender findings: 3) Tweets content

- In cooperation with University of Twente (Dong Nguyen)
- Machine learning approach that determines gender specific writing style
  - Language specific: Messages need to be Dutch!
  - 437 of 473 (92%) persons that created tweets could be classified

TweetGenie

Ik voorspel aan de hand van je tweets je leeftijd en geslacht.
Uitproberen? Vul dan hieronder je Twitter account in (je tweets moeten openbaar zijn).
Let op, ik begrijp alleen Nederlands! (only works for Dutch!)

pietdaas

Volgens mij is pietdaas een... man en circa 49 jaar oud.

Deel je resultaat op Twitter
Gender findings: 4) Profile picture

1) Face recognition
2) Standardisation of faces (resize & rotate)
3) Classify faces according to gender

- 603 of 804 (75%) profile pictures had 1 or more faces on it

- Use OpenCV to process pictures
Gender findings: overall results

<table>
<thead>
<tr>
<th></th>
<th>Diagnostic Odds Ratio (log)</th>
</tr>
</thead>
<tbody>
<tr>
<td>First name</td>
<td>4.33</td>
</tr>
<tr>
<td>Short bio</td>
<td>2.70</td>
</tr>
<tr>
<td>Tweet content</td>
<td>1.96</td>
</tr>
<tr>
<td>Picture (faces)</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Diagnostic Odds Ratio = \( \frac{TP}{FN} / \frac{FP}{TN} \)

random guessing

\[ \log(\text{DOR}) = 0 \]

- Multi-agent findings
  - Need ‘clever’ ways to combine these
  - Take processing efficiency of the ‘agent’ into consideration
Concluding remarks

- Social media is a difficult source to study
  - Contains a lot of ‘noise’
- Social media is a secondary data source
  - Produced for a ‘reason’ not identical to the one we want to use it for
    • A paradigm shift is needed (need a different mindset)
    • Try to improve quality (reduce noise; apply filter)
    • Make use of the large volume of data available
- Analysing texts and pictures is different/difficult
  - Learn by doing and by cooperating with experts
- Social media produces interesting results but
  - It is a relatively new area for official statistics, so a lot needs to be checked (this takes time)
- There are certainly possibilities for official statistics but
  - Is everybody in the office ready?
The Future

The future of statistics looks big.
Thank you for your attention!

@pietdaas