**Language reveals a lot about people**

Although social media are widely studied, computational linguistics typically focuses on prediction tasks:

- sentiment analysis
- authorship attribution
- personality prediction

Language analysis in social media can also be used to gain psychological insight.

**This work . . .**

. . . explores language features in Facebook as a function of gender, age, and personality:

- 74,941 volunteers shared their gender and age, and took a personality questionnaire
- 14.3m Facebook status updates resulting in 452m instances of language features (each volunteer had written at least 1000 words across their status updates)
- find language features most predictive of outcomes

correlations via multivariate linear regression allow for controls with other variables (i.e. correlations with gender, adjusted for age).

**Personality**

The well-accepted “Big Five” model (McCrae and John 1992):

- extraversion: active, assertive, energetic, enthusiastic, outgoing
- agreeableness: appreciative, forgiving, generous, kind
- conscientiousness: efficient, organized, planful, reliable
- neuroticism: anxious, self-pitying, tense, touchy, unstable
- openness: artistic, curious, imaginative, insightful, original

biopsychosocial characteristics that uniquely define a person (Friedman 2007).

**Features**

n-grams. \(j\) to \(l\) token sequences

- emoticon-aware tokenization
- stored as relative frequency
- collocation filter: \(pmi(\text{gram}) = \log \frac{\frac{\text{gram}}{\text{tok} \cdot \text{tok}}} {\frac{\text{tok}}{\text{tok}}}\)

topics. semantically-related words derived via LDA

- Latent Dirichlet Allocation (LDA): MALLET implementation (McCallum 2002)
- Adjusted hyper-parameters to favor fewer topics per document
- 2000 topics (tried 100, 500, 2000, 5000)
- usage per person: \(p(\text{topic}, \text{person}) = \sum_{\text{tok} \in \text{tok}} p(\text{topic|tok}) \cdot p(\text{tok|person})\)

**Results**

N-grams most distinguishing females (top) and males (bottom), adjusted for age: 
\(N = 74,941\); 46,572 females and 28,369 males; Bonferroni-corrected \(p < 0.001\).

N-grams and topics most distinguishing volunteers aged 13 to 18 and 23 to 29 
\(N = 74,941\); correlations adjusted for gender; Bonferroni-corrected \(p < 0.001\)

**Conclusions**

- A case-study on analyzing language in social media for psychological insight:
  - some results were known or obvious:
    - extraverts mention ‘party’
    - neuroticism and ‘depressed’
  - other revealed psychological insight:
    - ‘emotionally stable’ individuals mention more sports and life activities
    - older individuals mention more social topics and less anti-social topics
    - men preface ‘wife’ or ‘girlfriend’ with the possessive ‘my’ more often than woman do for ‘husband’ or ‘boyfriend’
- More sophisticated language analyses could be brought to bear.
  - features based on entity recognition or semantic relations
  - analyses which capture interactions between variables

**Age Plots**

Standardized frequency of topics and words across age. A. The best topic for each of the 4 age groups. B. Select social topics. C. ‘I’ and ‘we’ unigrams.

**Visualization**

1) Linguistic feature extraction
2) Correlation analysis
3) Visualization

Volunteer Data

- social media messages
- gender
- personality

- n-grams
- topics

• Adjusted hyper-parameters to favor fewer topics per document
• collocation filter:
•...