Toward Personality Insights from Language Exploration in Social Media

wwbp.org

H. Andrew Schwartz[†*], Johannes C. Eichstaedt[†], Lukasz Dziurzynski[†], Margaret L. Kern[†], Eduardo Blanco[‡], Michal Kosinski[ʃ], David Stillwell[ʃ], Martin E. P. Seligman[†], and Lyle H. Ungar[†]

[†]University of Pennsylvania [‡]Lymba Corporation [ʃ]University of Cambridge

*hansens@seas.upenn.edu

Language reveals a lot about people

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

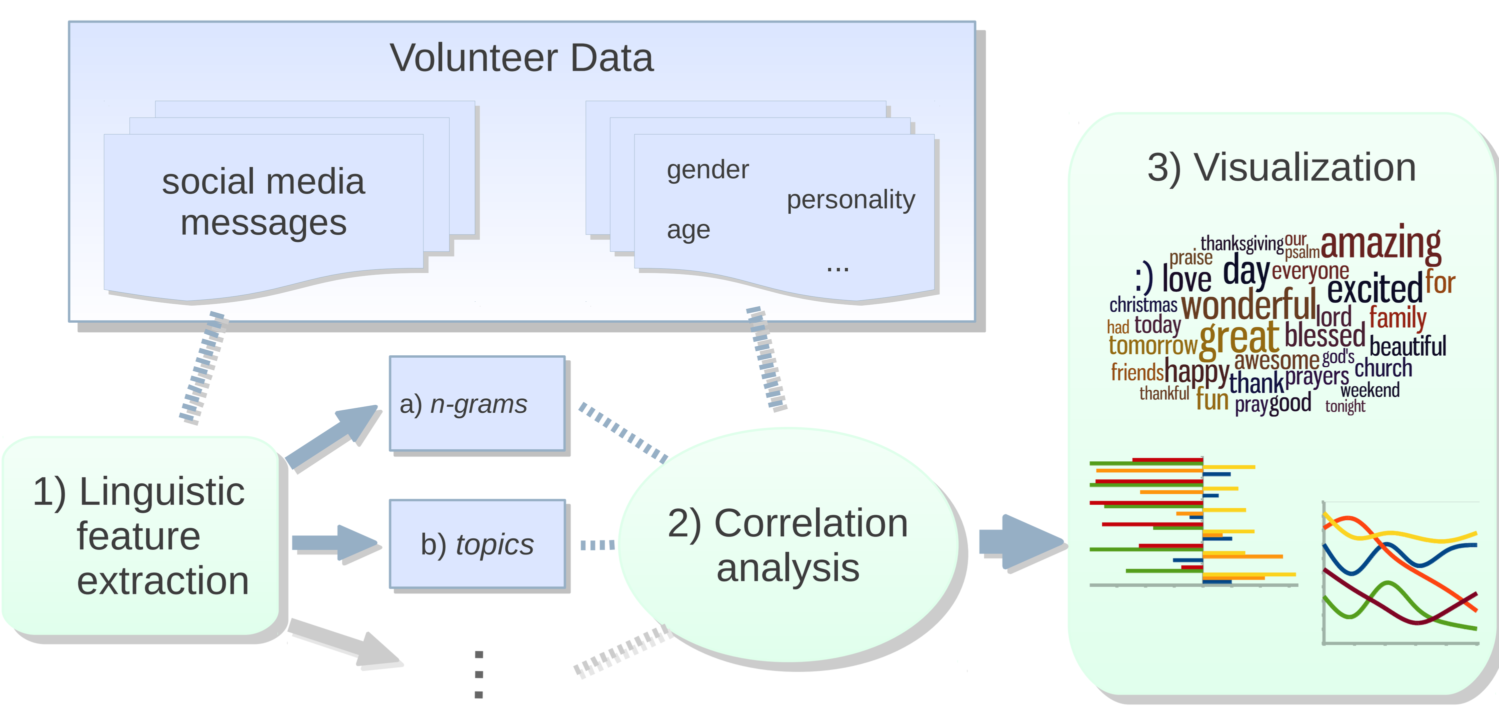
- sentiment analysis
- authorship attribution
- personality prediction
- ...

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-pittying, tense, touchy, unstable
- *openness*: artistic, curious, imaginative, insightful, original

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

Features

n-grams. *1 to 3 token sequences*

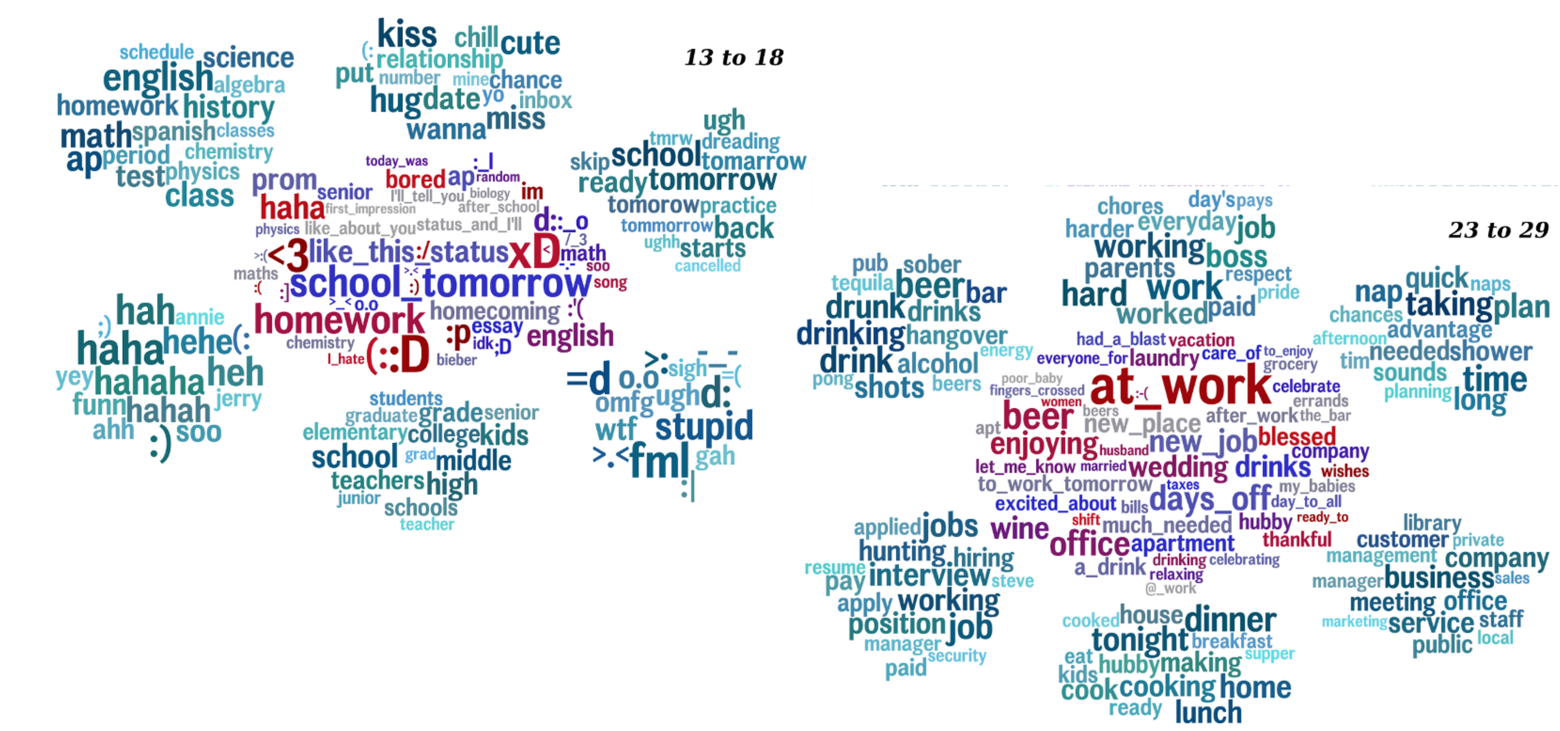
- emoticon-aware tokenization
- stored as relative frequency
- collocation filter: $pmi(ngram) = \log \frac{p(ngram)}{\prod_{token \in ngram} p(token)}$

topics. *semantically-related words derived via LDA*

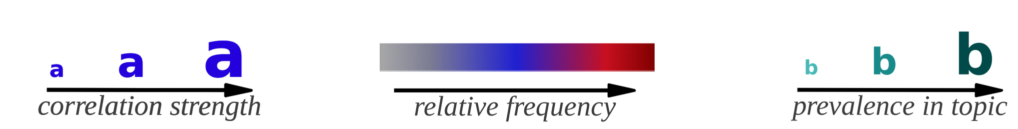
- Latent Dirichlet Allocation (LDA); MALLET implementation (McCallum 2002)
- Adjusted hyper-parameters to favor fewer topics per document
- 2000 topics (tried 100, 500, 2000, 5000)
- usage per person: $p(topic, person) = \sum_{tok \in topic} p(topic|tok) * p(tok|person)$

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 $p < 0.001$)



N-grams most distinguishing *extraversion* from *introversion* and *neuroticism* from *emotional stability* ($N = 72,791$ for extraversion; $N = 72,047$ for neuroticism; adjusted for age and gender; Bonferroni-corrected $p < 0.001$).

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.

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 prefer mention '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