Toward Insight

- Growth of social media => unprecedented amounts of personal discourse
- Leveraged in computational linguistics mostly for prediction:
  - sentiment analysis
  - authorship attribution
  - personality prediction

Toward Insight

- Growth of social media => unprecedented amounts of personal discourse
- Leveraged in computational linguistics mostly for prediction:
  - sentiment analysis
  - authorship attribution
  - personality prediction

- Analysis of language in social media can also be used for the purpose of gaining psychological insights.
This Work

- Explores language in social media as a function of gender, age, and personality.

- Data Set over Facebook
  - ~75,000 volunteers, writing 14m status updates
    - wrote at least 1000 words
    - reported gender and age
    - took standard personality survey: “the big five” / five factor model.
      (MyPersonality App.)
Method

Differential Language Analysis

Volunteer Data

social media messages
gender
age
personality

1) Linguistic feature extraction

a) n-grams
b) topics

2) Correlation analysis

3) Visualization

wwwbp.org
Features

• N-grams
  – emoticon-aware tokenization
  – collocation filter based on point-wise mutual information; used by at least 1% of volunteers

• LDA topics
  – Latent Dirichlet Allocation
  – 2000 topics from larger 20m status updates
  – usage defined as:

\[
p(topic, person) = \sum_{tok \in topic} p(topic|tok) \times p(tok|person)
\]
Explicit Language Warning
Results: Gender

Bonferroni-adjusted p < 0.001
Results: Personality - *Extraversion*
Results: Personality - *Introversion*
Results: Age

23 to 29
Conclusions

- This was a case study in gaining personality insights
  - expected results
  - insightful results
- Moving Forward
  - more feature types (semantic relations, named entities)
  - more analyses (capture interactions)
Conclusions

• This was a case study in gaining personality insights
  – expected results
  – insightful results

• Moving Forward
  – more feature types (semantic relations, named entities)
  – more analyses (capture interactions)

• Other Works
  – predicting gender (91.9%), age ($r = 0.84$), personality ($r = 0.31$ to $0.42$)
  – characterizing geographic variance in well-being and health
Personality

• “The Big Five” / Five Factor 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