웹 로그인 아이디와 사용자 성별 및 연령대 관계 분석과 예측 모형 - 관련 연구

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9), e73791.

2015. 11. 25.

박희웅

서울대학교 산업공학과

The paper is

- Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach

- Written by H. Andrew Schwartz[1,2*], Johannes C. Eichstaedt[1], Margaret L. Kern[1], Lukasz Dziurzynski[1], Stephanie M. Ramones[1], Megha Agrawal[1,2], Achal Shah[2], Michal Kosinski[3], David Stillwell[3], Martin E. P. Seligman[1], Lyle H. Ungar[2]
  1. Positive Psychology Center, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
  2. Computer & Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America
  3. The Psychometrics Centre, University of Cambridge, Cambridge, United Kingdom

- Supported by the Robert Wood Johnson Foundation's Pioneer Portfolio, through a grant to Martin Seligman, "Exploring Concept of Positive Health"

Contributions

- Largest study of personality and language use
  - a data-driven collection of words, phrases, and topics, in which the lexicon is based on the words of the text being analyzed
- Further insights into the behavioral residue of personality types
- Informative features via their use in predictive model
  - accuracy of 91.9% for gender prediction
- Word cloud visualization
- Providing comprehensive word, phrase, and topic correlation data

Closed vs Open vocabulary

- Closed vocabulary: word-category lexica

$$p(category|subject) = \frac{\sum_{word \in category} freq(word, subject)}{\sum_{word \in vocab(subject)} freq(word, subject)}$$

  - Least square regression to link word categories with author attribute
    - ✓ Category Linguistic Inquiry and Word Count (LIWC) as category

      LIWC includes 64 different categories of language from part-of-speech to topical categories

- Open vocabulary: Differential Language Analysis
  - Not limited to predefined word lists (i.e. data-driven)
  - Discriminating features selected by using significance tests

Open vocabulary

- Linguistic feature extraction
  - Words and phrases
    - ✓ Consisted of sequences of 1 to 3 words
    - ✓ Including emoticons
    - ✓ When extracting phrases, keep only phrases with high information value
    - ✓ Keep only words used by at least 1% of subjects
  - Topics
    - ✓ Latent Dirichlet Allocation
    - ✓ 2000 topics extracted

0	1	2	3	4	5	6	7	8	9
turtle	asked	hate	wanna	snow	car	forward	weekend	chocolate	cool
giant	told	haters	hang	roads	cop	lookin	awesome	eating	pretty
shell	wanted	hating	text	drive	cars	ward	amazing	hot	nice
ninja	replied	jealous	bored	car	parking	marcus	great	milk	guy
turtles	boy	hater	call	driveway	ticket	foward	retreat	icecream	super
penguin	looked	bitches	talk	ice	driving	seein	couples	covered	hang

Open vocabulary

- Correlation analysis
  - Ordinary least square analysis
  - Age and gender as covariates
  - The coefficient of the target explanatory variable as its correlation
  - Bonferroni-corrected significance test

- Visualization
  - Word size according to the strength of the correlation of the word
  - Color to represent frequency
  - Six most distinguishing topics on the perimeter of the word clouds

Dataset

- Facebook status updates
  - 19 million updates written by 136,000 volunteered participants
  - Participants took 20 to 100 questions about personality.
  - Restrict to users
    - ✓ Speaking English as a primary language
    - ✓ Writing At least a thousand words
    - ✓ Less than 65 years
    - ✓ Indicating Gender and age
  - As a result, 74,941 volunteers, writing a total of 309 million words across 15.4 million status updates
    - ✓ 4,129 words over 206 status updates per person

	N	mean	std.	skewness
Gender	74859	0.62	0.49	-0.5
Age	74859	23.43	8.96	1.77
Extraversion	72709	-0.07	1.01	-0.37
Agreeableness	72772	0.03	1	-0.4
Conscientiousness	72781	-0.04	1.01	-0.09
Neuroticism	71968	0.14	1.04	-0.21
Openness	72809	0.12	0.97	-0.48

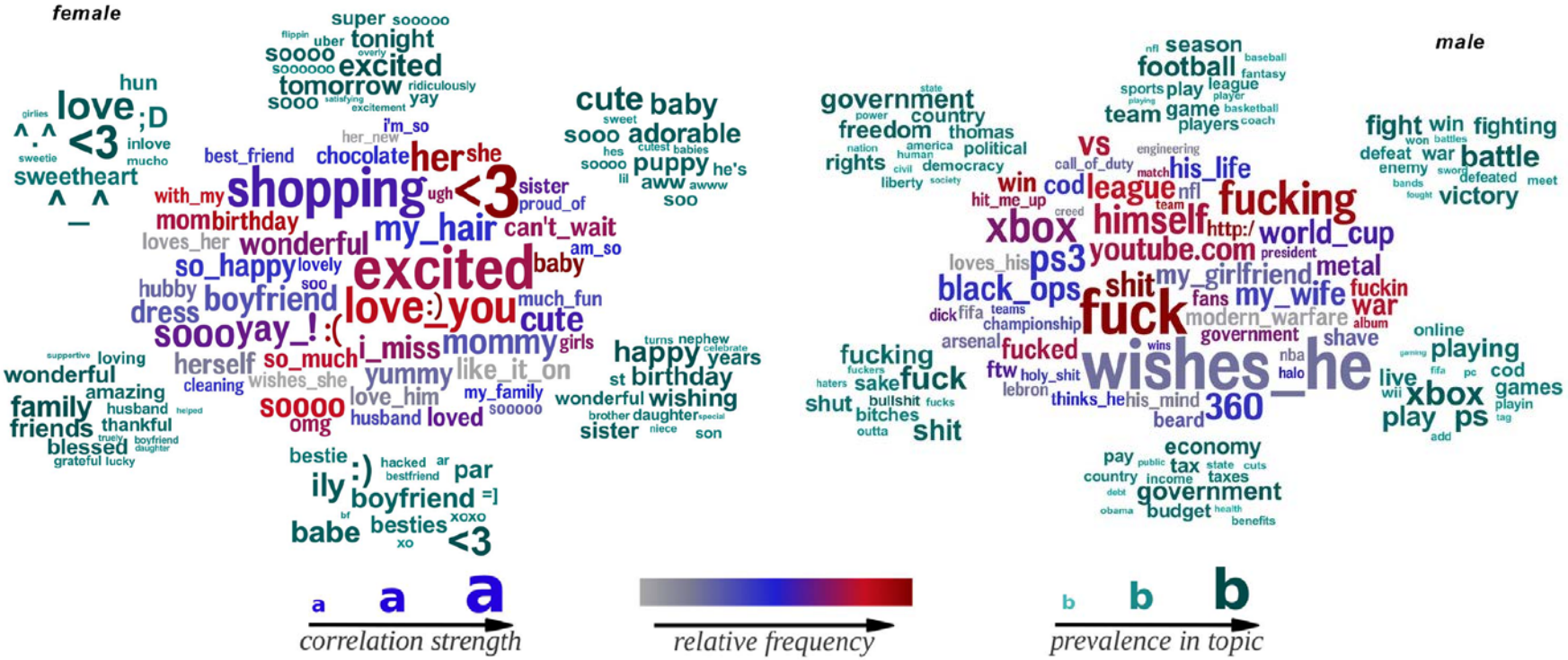
Results of closed vocabulary

- Correlation values of LIWC categories

LIWC Category	Gender		Age		Extraversion		Agreeableness		Conscientious.		Neuroticism		Openness	
	[34] d	our β	[30] β	our β	[27] ρ	our β	[27] ρ	our β	[27] ρ	our β	[27] ρ	our β	[27] ρ	our β
Total function words	-	-0.04	-	0.16	-	-0.04	-	0.02	-	0.02	-	0.03	-	0.09
Total pronouns	0.36	0.07	-	-0.02	*ns*	*ns*	0.11	*ns*	*ns*	-0.03	*ns*	0.04	-0.21	0.07
Personal pronouns	-	0.14	-	-0.08	-	*ns*	-	*ns*	-	-0.04	-	0.04	-	0.05
1st pers singular	0.17	0.13	-0.14	-0.22	*ns*	*ns*	*ns*	-0.03	*ns*	-0.06	0.12	0.05	-0.16	0.05
1st pers plural	*ns*	*ns*	-0.13	0.21	0.11	0.03	0.18	0.05	*ns*	0.05	*ns*	-0.04	-0.1	*ns*
2nd person	-0.06	0.05	-	0.04	0.16	*ns*	*ns*	0.02	*ns*	*ns*	-0.15	*ns*	-0.12	0.02
3rd pers singular	-	0.09	-	0.15	-	*ns*	-	*ns*	-	*ns*	-	0.02	-	*ns*
3rd pers plural	-	-0.05	-	0.26	-	-0.06	-	-0.04	-	*ns*	-	0.02	-	0.03
3rd pers overall	0.2	-	-	-	*ns*	-	*ns*	-	*ns*	-	*ns*	-	*ns*	-
Impersonal pronouns	-	-0.09	-	0.11	-	-0.05	-	*ns*	-	*ns*	-	0.02	-	0.08
Articles	-0.24	-0.24	-	0.28	*ns*	-0.05	*ns*	*ns*	0.09	0.02	-0.11	-0.02	0.2	0.13
Common verbs	-	0.04	-	0.02	-	-0.03	-	*ns*	-	*ns*	-	0.04	-	0.03
Auxiliary verbs	-	0.02	-	0.08	-	-0.06	-	*ns*	-	*ns*	-	0.05	-	0.07
Past tense	0.12	-0.03	-0.16	*ns*	*ns*	-0.04	0.1	0.02	*ns*	-0.02	*ns*	*ns*	-0.16	*ns*
Present tense	0.18	0.08	0.04	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	*ns*	0.04	-0.16	0.03
Future tense	*ns*	-0.07	0.14	0.09	*ns*	-0.05	*ns*	*ns*	*ns*	*ns*	*ns*	0.03	*ns*	0.05
Adverbs	-	0.05	-	-0.07	-	-0.04	-	*ns*	-	*ns*	-	0.05	-	0.04
Prepositions	-0.17	-0.13	-	0.27	*ns*	-0.04	*ns*	0.03	*ns*	0.06	*ns*	*ns*	0.17	0.06
Conjunctions	-	0.03	-	0.12	-	-0.02	-	0.02	-	0.02	-	0.02	-	0.06
Negations	0.11	*ns*	-	-0.12	*ns*	-0.06	*ns*	-0.05	-0.17	-0.03	0.11	0.07	-0.13	0.02
Quantifiers	-	-0.09	-	0.24	-	-0.02	-	0.03	-	0.05	-	*ns*	-	0.05
Numbers	-0.15	-0.13	-	0.05	-0.12	-0.06	0.11	0.02	*ns*	0.02	*ns*	*ns*	-0.08	0.06
Swear words	-0.22	-0.21	-	-0.17	*ns*	*ns*	-0.21	-0.15	-0.14	-0.09	0.11	0.06	*ns*	*ns*
Social processes	-	0.08	-0.13	0.21	0.15	0.04	0.13	0.02	*ns*	*ns*	*ns*	*ns*	-0.14	*ns*
Family	0.12	0.22	-	0.28	0.09	0.03	0.19	0.03	*ns*	0.03	*ns*	*ns*	-0.17	-0.12
Friends	0.09	0.08	-	0.26	0.15	0.05	0.11	0.04	*ns*	0.02	-0.08	*ns*	*ns*	-0.04
Humans	*ns*	0.04	-	0.06	0.13	0.06	*ns*	-0.05	-0.12	*ns*	*ns*	*ns*	-0.09	*ns*
Affective processes	0.11	0.11	-	-0.05	0.09	0.07	*ns*	0.02	*ns*	*ns*	*ns*	*ns*	-0.12	-0.04
Positive emotion	*ns*	0.21	0.12	0.14	0.1	0.13	0.18	0.13	*ns*	0.1	*ns*	-0.08	-0.15	-0.07
Negative emotion	0.1	-0.12	-0.05	-0.31	*ns*	-0.07	-0.15	-0.17	-0.18	-0.13	0.16	0.15	*ns*	0.03
Anxiety	0.16	0.08	-	-0.13	*ns*	-0.04	*ns*	-0.02	*ns*	-0.02	0.17	0.06	*ns*	0.07

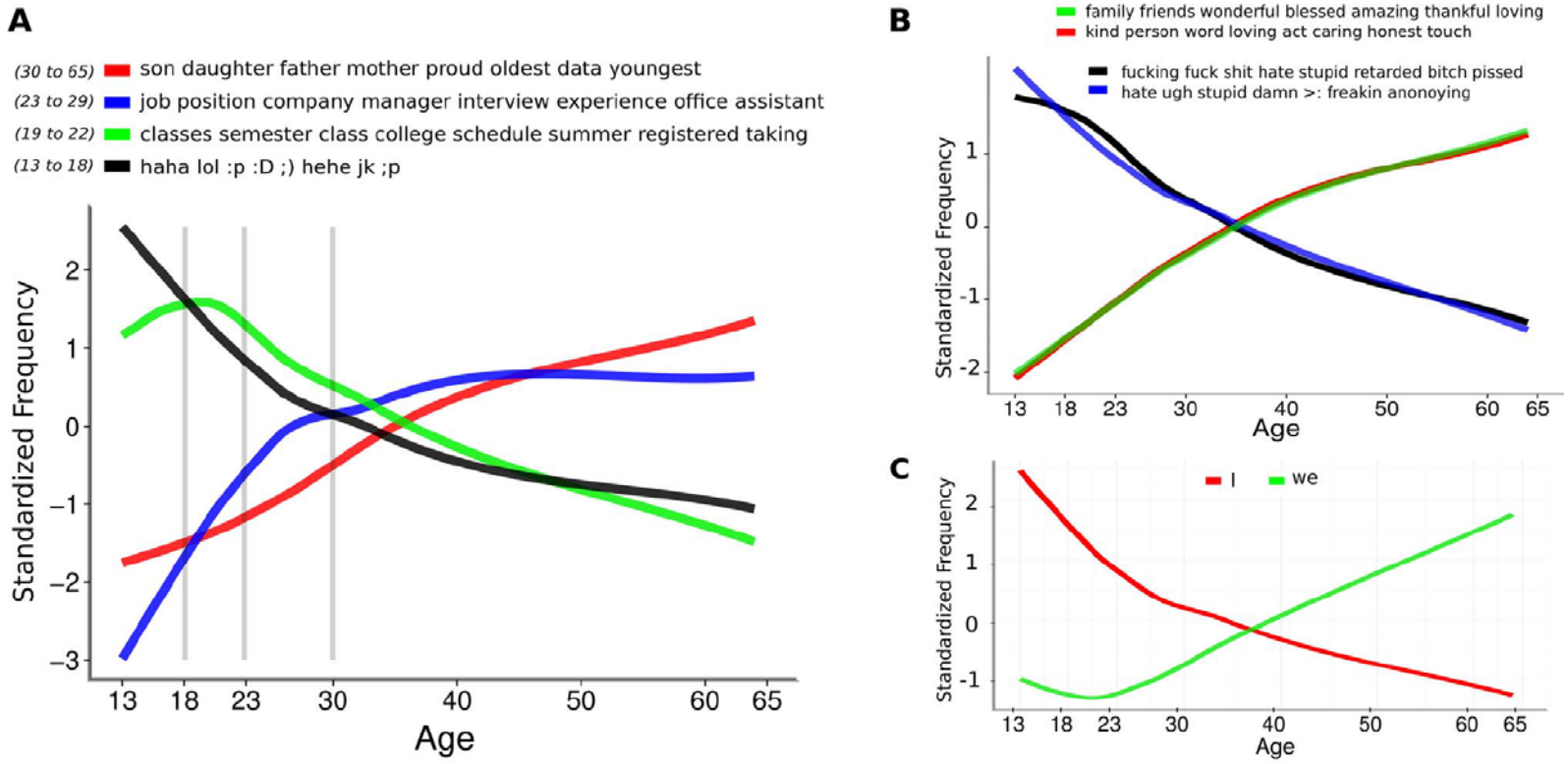
Results of open vocabulary

- Most discriminating words and topics



Results of open vocabulary

- Frequency of words and topics



Prediction evaluation

- SVM for gender
- Ridge regression for age and each factor of personality
- Training set 75%, test set 25%, 10% of training set used for parameter setting

features	Gender accuracy	Age R	Extraversion R	Agreeableness R	Conscientious R	Neuroticism R	Openness R
LIWC	78.40%	0.65	0.27	0.25	0.29	0.21	0.29
Topics	87.50%	0.8	0.32	0.29	0.33	0.28	0.38
WordPhrases	91.40%	0.83	0.37	0.29	0.34	0.29	0.41
WordPhrases + Topics	91.90%	0.84	0.38	0.31	0.35	0.31	0.42
Topics + LIWC	89.20%	0.8	0.33	0.29	0.33	0.28	0.38
WordPhrases + LIWC	91.60%	0.83	0.38	0.3	0.34	0.3	0.41
WordPhrases + Topics + LIWC	91.90%	0.84	0.38	0.31	0.35	0.31	0.42

Outline

- 서론
- 아이디의 특성
- HMM을 이용한 아이디 구조 분석 방법
- 구조 분석 결과
- 결론