Developing Age and Gender Predictive Lexica over Social Media

Maarten Sap[1] **Gregory Park**[1] **Johannes C. Eichstaedt**[1] **Margaret L. Kern**[1]
David Stillwell[3] **Michal Kosinski**[3] **Lyle H. Ungar**[2] and **H. Andrew Schwartz**[2]

[1]Department of Psychology, University of Pennsylvania
[2]Computer & Information Science, University of Pennsylvania
[3]Psychometrics Centre, University of Cambridge
maarten@sas.upenn.edu

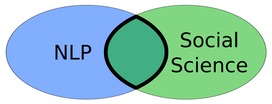
wwbp.org

Lexica in Social Sciences

Lexica are widely accepted and used in social sciences

- Most of them are manually curated - word by word
- They are easily used, and make a lot of sense for social scientists
- LIWC (Pennebaker et al., 2001) had over 1,000 citations in 2013 only

Goal: *as accurate as modern NLP techniques & as accessible as widely used social science lexica.*



- Introducing weighted lexica as an alternative format for machine learning models
- Using a bottom-up approach to generate the lexica, as opposed to a top-down, manual one

Age and Gender on Social Media

Language, behavior and health correlate with age and gender

- Women tend to live longer (CDC, 2014); people, with age, tend to be more agreeable, more conscientious and less open to experience (McCrae et al. 1999)
- most social scientific studies on social media data use biased samples of age and gender

There is a need for accessible tools to predict demographic variables for social science, economic, and business applications.

Data

75,394 Facebook users from the MyPersonality app (Kosinski and Stillwell, 2012):

- **main test set** - randomly sampled 1,000 users
- **stratified test set** - equal proportions of 1,520 males and females across 12 four-year age bins (age 13-60), independent from the main test set
- **training set** - remaining 72,874 users

15,006 Blogger users from Schler et al. (2006):

- **test set** - randomly sampled 1,000 users
- **training set** - remaining 15,006 users

11,000 Twitter users, random sample from Volkova et al. (2013)

- **test set** - randomly sampled 1,000 users
- **training set** - remaining 10,000 users

All data sets contained users that had written a minimum of 1000 words, with age information. Gender was included for Facebook users and Bloggers.

Lexicon Creation

Goal: Simple yet effective method

Lexicon Extraction:

$$usage_{lex} = \sum_{word \in lex} w_{lex}(word) * \frac{freq(word, doc)}{freq(*, doc)}$$

where $w_{lex}(word)$: lexicon (lex) weight for the $word$, $freq(word, doc)$: frequency of the word in the document (or for a given user), and $freq(*, doc)$: total word count for that document (or user).

Linear Multivariate Regression:

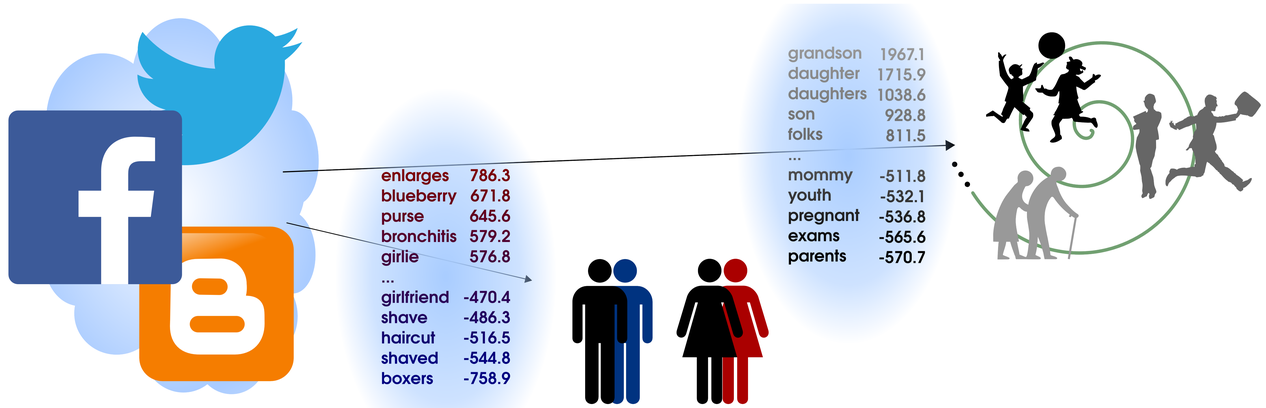
$$y = (\sum_{f \in features} w_f * x_f) + w_0$$

where x_f is the value for a feature (f), w_f is the feature coefficient, and w_0 is the intercept (a constant fit to shift the data such that it passes through the origin).

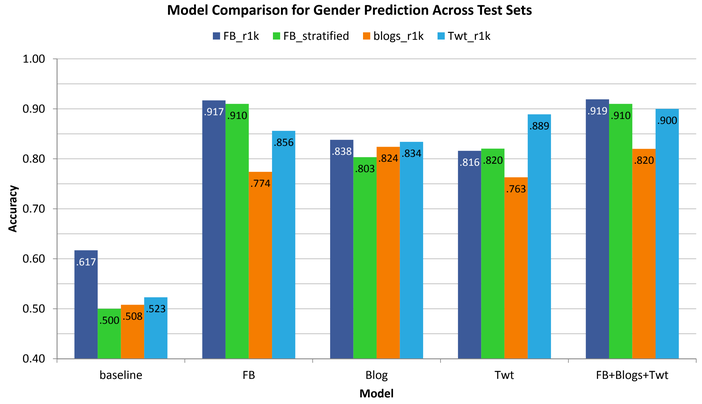
In the case of regression, y is the outcome value (e.g. age) while in classification, y is used to separate the classes (e.g. $>= 0$ is female, < 0 is male).

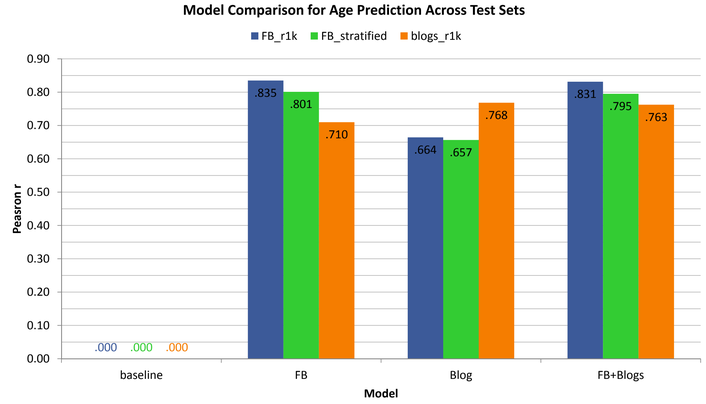
. . . multivariate modeling can be seen as learning a weighted lexicon and an intercept

Predicting Demographics from Social Media

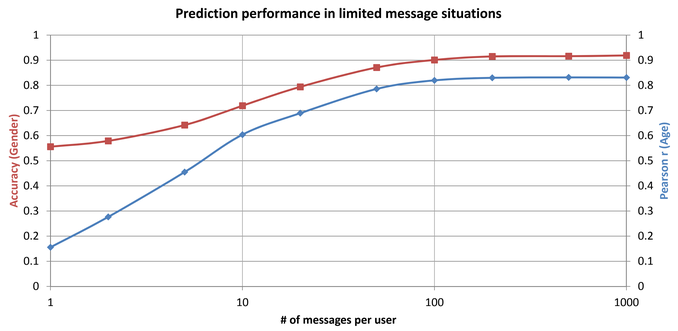


Prediction Performance of Lexica





Limited messages



Conclusion

Easy to use gender and age lexica are now available that . . .

- are in line with state-of-the-art accuracies for age ($r = .831$) and gender (91.9% accuracy)
- have predictive power that generalizes accross multiple social media platforms
- maintains reasonable accuracies when the number of messages per user is limited

. . . can be downloaded from wwbp.org/data.html, along with instructions for use

Given that manual lexica are already extensively employed in social sciences such as psychology, economics, and business, using lexical representations of data-driven models allows the utility of our models to extend beyond the borders of the field of NLP.

download @
wwbp.org/data.html