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

Abstract

Demographic lexica have potential for widespread use in social science, economic, and business applications. We derive predictive lexica (words and weights) for age and gender using regression and classification models from word usage in Facebook, blog, and Twitter data with associated demographic labels. The lexica, made publicly available,[1] achieved state-of-the-art accuracy in language based age and gender prediction over Facebook and Twitter, and were evaluated for generalization across social media genres as well as in limited message situations.

1 Introduction

Use of social media has enabled the study of psychological and social questions at an unprecedented scale (Lazer et al., 2009). This allows more data-driven discovery alongside the typical hypothesis-testing social science process (Schwartz et al., 2013b). Social media may track disease rates (Paul and Dredze, 2011; Google, 2014), psychological well-being (Dodds et al., 2011; De Choudhury et al., 2013; Schwartz et al., 2013a), and a host of other behavioral, psychological and medical phenomena (Kosinski et al., 2013).

Unlike traditional hypothesis-driven social science, such large-scale social media studies rarely take into account—or have access to—age and gender information, which can have a major impact on many questions. For example, females live almost five years longer than males (cdc, 2014; Marengoni et al., 2011). Men and women, on average, differ markedly in their interests and work preferences (Su et al., 2009). With age, personalities gradually change, typically becoming less open to experiences but more agreeable and conscientious (McCrae et al., 1999). Additionally, social media language varies by age (Kern et al., 2014; Pennebaker and Stone, 2003) and gender (Huffaker and Calvert, 2005). Twitter may have a male bias (Mislove et al., 2011), while social media in general skew towards being young and female (pew, 2014).

Accessible tools to predict demographic variables can substantially enhance social media's utility for social science, economic, and business applications. For example, one can post-stratify population-level results to reflect a representative sample, understand variation across age and gender groups, or produce personalized marketing, services, and sentiment recommendations; a movie may be generally disliked, except by people in a certain age group, whereas a product might be used primarily by one gender.

This paper describes the creation of age and gender predictive lexica from a dataset of Facebook users who agreed to share their status updates and reported their age and gender. The lexica, in the form of words with associated weights, are derived from a penalized linear regression (for continuous valued age) and support vector classification (for binary-valued gender). In this modality, the lexica are simply a transparent and portable means for distributing predictive models based on words. We test generalization and adapt the lexica to blogs and Twitter, plus consider situations when limited messages are available. In addition to use in the computational linguistics community, we believe the lexicon format will make it easier for social scientists to leverage data-driven models where manually created lexica currently dominate[2] (Dodds et al., 2011; Tausczik and Pennebaker, 2010).

2 Related Work

Online behavior is representative of many aspects of a user's demographics (Pennacchiotti and Popescu, 2011; Rao et al., 2010). Many studies have used linguistic cues (such as ngrams) to determine if someone belongs to a certain age group, be it on Twitter or another social media platform (Al Zamal et al., 2012; Argamon et al., 2009; Nguyen et al., 2013; Rangel and Rosso, 2013). Gender prediction has been studied across blogs (Burger and Henderson, 2006; Goswami et al., 2009), Yahoo! search queries (Jones et al., 2007), and Twitter (Burger et al., 2011; Nguyen et al., 2013; Liu and Ruths, 2013; Rao et al., 2010). Because Twitter does not make gender or age available, such work infers gender and age by leveraging profile information, such as gender-discriminating names or crawling for links to publicly available data (e.g. Burger et al.,

[1]download at http://www.wwbp.org/data.html

[2]The LIWC lexicon, derived manually based on psychological theory, (Pennebaker et al., 2001) had 1136 citations in 2013 alone.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1146–1151,
October 25-29, 2014, Doha, Qatar.

2011).

While many studies have examined prediction of age or gender, none (to our knowledge) have released a model to the public, much less in the form of a lexicon. Additionally, most works in age prediction classify users into bins rather than predicting a continuous real-valued age as we do (exceptions: Nguyen et al., 2013; Jones et al., 2007). People have also used online media to infer other demographic-like attributes such as native language (Argamon et al., 2009), origin (Rao et al., 2010), and location (Jones et al., 2007). An approach similar to the one presented here could be used to create lexica for any of these outcomes.

While lexica are not often used for demographics, data-driven lexicon creation over social media has been well studied for sentiment, in which univariate techniques (e.g. *point-wise mutual information*) dominate[3]. For example, Taboada et al. (2011) expanded an initial lexicon by adding on co-occurring words. More recently, Mohammad's sentiment lexicon (Mohammad et al., 2013) was found to be the most informative feature for the top system in the SemEval-2013 social media sentiment analysis task (Wilson et al., 2013). Approaches like point-wise mutual information take a univariate view on words–i.e. the weight given to one feature (word) is not affected by other features. Since language is highly collinear, we take a multivariate lexicon development approach, which takes covariance into account (e.g. someone who mentions 'hair' often is more likely to mention 'brushing', 'style', and 'cut'; weighting these words in isolation might "double-count" some information).

3 Method

Primary data. Our primary dataset consists of Facebook messages from users of the MyPersonality application (Kosinski and Stillwell, 2012). Messages were posted between January 2009 and October 2011. We restrict our analysis to those Facebook users meeting certain criteria: they must indicate English as a primary language, have written at least 1,000 words in their status updates, be younger than 65 years old (data beyond this age becomes very sparse), and indicate their gender and age. This resulted in a dataset of N = 75,394 users, who wrote over 300 million words collectively. We split our sample into training and test sets. Our primary *test set* consists of a 1,000 randomly selected Facebook users, while the *training set* that we used for creating the lexica was a subset (N = 72,874) of the remaining users.

Additional data To evaluate our predictive lexica in differing situations, we utilize three additional datasets: stratified Facebook data, blogs, and tweets. The stratified Facebook data (exclusively used for testing) consists of equal proportions of 1,520 males and females across 12 4-year age bins starting at 13 and ending at 60.[4] This roughly matchs the size of the main *test set*.

Seeking out-of-domain data, we downloaded age and gender annotated blogs from 2004 (Schler et al., 2006) (also used in Goswami et al., 2009) and gender labeled tweets (Volkova et al., 2013). Limiting the sample to users who wrote at least 1000 words, the total number of bloggers is 15,006, of which 50.6% are female and only 15% are over 27 (reflecting the younger population standard in social media). From this we use a randomly selected 1,000 bloggers as a blogger test set and the remaining 14,006 bloggers for training. Similarly for the Twitter dataset, we use 11,000 random gender-only annotated users, in which 51.9% are female. We again randomly select 1,000 users as a test set for gender prediction and use the remaining 10,000 for training.

3.1 Lexicon Creation

We present a method of weighted lexicon creation by using the coefficients from linear multivariate regression and classification models. Before delving into the creation process, consider that a weighted lexicon is often applied as the sum of all weighted word relative frequencies over a document:

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

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

Further consider how one applies linear multivariate models in which the goal is to optimize feature coefficients that best fit the continuous outcome (regression) or separate two classes (classification):

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

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 classes (e.g. $>= 0$ is female, < 0 is male). If all features are word relative frequencies ($\frac{freq(word,doc)}{freq(*,doc)}$) then many multivariate modeling techniques can simply be seen as learning a weighted lexicon plus an intercept[5].

[3]Note that the point-wise information-derived sentiment lexica are often used as features in a supervised model, essentially dimensionally reducing a large set of words into positive and negative sentiment, while our lexica represent the predictive model itself.

[4]65 females and 65 males in each of the first 11 bins: [13,16], [17,20], ..., [53, 56]; the last bin ([57, 60]) contained 45 males and 45 females. The [61.64] bin was excluded as it was much smaller.

[5]included in the lexicon distribution

	age						gender			
model\corpus	**randFB**		**stratFB**		**randBG**		**randFB**	**stratFB**	**randBG**	**randT**
	r	mae	r	mae	r	mae	acc	acc	acc	acc
baseline	0	6.14	0	11.62	0	6.11	.617	.500	.508	.518
FB_{lex}	.835	3.40	.801	6.94	.710	5.76	.917	.913	.774	.856
BG_{lex}	.664	4.26	.656	11.39	.768	3.63	.838	.803	.824	.834
FB+BG_{lex}	.831	3.42	.795	7.06	.762	3.76	.913	.909	.822	.858
T_{lex}							.816	.820	.763	.889
FB+BG+T_{lex}							.919	.910	.820	.900

Table 1: Prediction accuracies for age (Pearson correlation coefficient(r); mean absolute error (mae) in years) and gender (accuracy %). Baseline for age is mean age of training sample; for gender, it is the most frequent class (female). Lexica tested include those derived from Facebook (FB_{lex}), blogs (BG_{lex}), and Twitter (T_{lex}). We evaluate over a random Facebook sample (randFB), a stratified Facebook sample (stratFB), a random blogger sample (randBG), and a random twitter sample (randT). All results were a significant ($p < 0.001$) improvement over the baseline.

In practice, we learn our 1gram coefficients (i.e. lexicon weights) from ridge regression (Hoerl and Kennard, 1970) for age (continuous variable) and from support vector classification (Fan et al., 2008) for gender (binary variable). Ridge regression uses an L2 ($\alpha||\beta||^2$) penalization to avoid overfitting (Hoerl and Kennard, 1970). Although some words no doubt have a non-linear relationship with age (e.g., 'fiance' peaks in the 20s), we still find high accuracy from a linear model (see Table 1) and it allows for a distribution of the model in the accessible form of a lexicon. For gender prediction, we use an SVM with a linear kernel with L1 penalization ($\alpha||\beta||_1$) (Tibshirani, 1996). Because the L1 penalization zeros-out many coefficients, it has the added advantage of effectively reducing the size of the lexica. Using the training data, we test a variety algorithms including the lasso, elastic net regression, and L2 penalized SVMs in order to decide which learning algorithms to use.

To extract the words (1grams) to use as features and which make up lexica, we use the Happier Fun Tokenizer,[6] which handles social media content and markup such as emoticons or hashtags. For our main user-level models, word usage is aggregated as the relative frequency ($\frac{freq(word,user)}{freq(*,user)}$). Due to the sparse and large vocabulary of social media data, we limit the 1grams to those used by at least 1% of users.

4 Evaluation

We evaluate our predictive lexica across held-out user data. First, we see how well lexica derived from Facebook users predict a random set of additional users. Then, we explore generalization of the models in various other settings: on a stratified Facebook test sample, blogs, and Twitter. Finally, we compare lexica fit to a restricted number of messages per user.

Results of our evaluation over Facebook users are shown in Table 1 (randFB columns). Accuracies for age are reported as Pearson correlation coefficients (r) and mean absolute errors (mae), measured in years. For gender, we use an accuracy % (number-correct over test-size). As baselines, we use the mean for age (23.0 years old) and the most frequent class (female) for gender. We see that for both age and gender, accuracies are substantially higher than the baseline. These accuracies were just below with no significant difference previous state-of-the-art results (Schwartz et al., 2013; r = 0.84 for age and 91.9% accuracy for gender). [7]

Because of the nature of our datasets (the Facebook data is private) and task (*user-level* predictions), comparable previous studies are nearly nonexistent. Nonetheless, the Twitter data was a random subset of users based on the (Burger et al., 2011) dataset excluding non-English tweets, making it somewhat comparable. In this case, the lexica outperformed previous results for gender prediction of Twitter users, which ranged from 75.5% to 87% (Burger et al., 2011; Ciot et al., 2013; Liu and Ruths, 2013; Al Zamal et al., 2012). However, the lexica were unable to match the 92.0% accuracy Burger et al. (2011) achieved when using profile information in addition to text. No other similar studies — to the best of our knowledge — have been conducted.

Application in other settings. While Facebook is the ideal setting to apply our lexica, we hope that they generalize to other situations. To evaluate their utility in other settings, we first tested them over a gender and age stratified Facebook sample. Our random sample, like all of Facebook, is biased toward the young; this stratified test sample contains equal numbers of males and females, ages 13 to 60. Next, we use the lexica to predict data from other domains: blogs (Schler et al., 2006) and Twitter (Volkova et al., 2013). In this case, our goal was to account for the content and stylistic variation that may be specific to Facebook.

[6] downloaded from http://www.wwbp.org/data.html

[7] Adding 2 and 3-grams increases the performance of our model (r = 0.85, 92.7%), just above our previous results (Schwartz et al., 2013b). However, with the accessibility of single word lexica in mind, this current work focuses on features based entirely on 1grams.

# Msgs:	**all**	**100**	**20**	**5**	**1**
age	**.831**	.820	.688	.454	.156
gender	**.919**	.901	.796	.635	.554

Table 2: Prediction accuracies for age (Pearson correlation) and gender (accuracy %) when reducing the number of messages from each user.

Results over these additional datasets are shown in Table 1 (stratFB, randBG, and randT columns). The performance decreases as expected since these datasets have differing distributions, but it is still substantially above mean and most frequent class baselines on the stratified dataset. Over blogs and Twitter, both age and gender prediction accuracies drop to a greater degree (when only using the Facebook-trained models), suggesting stylistic or content differences between the domains. However, when using lexica created with data from across multiple domains, the results in Facebook, blogs, and Twitter remain in line with results from models created specifically over their respective domains. In light of this result, we release the FB+BG age & FB+BG+T gender models as lexica (available at www.wwbp.org/data.html).

Limiting messages per user. As previously noted, some applications of demographic estimation require predictions over more limited messages. We explore the accuracy of user-level age and gender predictions as the number of messages per user decreases in Table 2. For these tests we used the FB+BG age & FB+BG+T gender lexica. Confirming findings by Van Durme (2012), the fewer posts one has for each user, the less accurate the gender and age predictions. Still, given the average user posted 205 messages, it seems that not all messages from a user are necessary to make a decent inference on their age and gender. Future work may explore models developed specifically for these limited situations.

5 Conclusion

We created publicly available lexica (words and weights) using regression and classification models over language usage in social media. Evaluation of the lexica over Facebook yielded accuracies in line with state-of-the-art age ($r = 0.831$) and gender (91.9% accuracy) prediction. By deriving the lexica from Facebook, blogs, and Twitter, we found the predictive power generalized across all three domains with little sacrifice to any one domain, suggesting the lexica may be used in additional social media domains. We also found the lexica maintain reasonable accuracy when writing samples were somewhat small (e.g. 20 messages) but other approaches may be best when dealing with more limited data.

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.

Acknowledgement

Support for this work was provided by the Templeton Religion Trust and by Martin Seligman of the University of Pennsylvania's Positive Psychology Center.

References

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

John D Burger and John C Henderson. 2006. An exploration of observable features related to blogger age. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 15–20.

John D Burger, John C Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

2014. Faststats: How healthy are we. `http://www.cdc.gov/nchs/fastats/healthy.htm`. Accessed on March 12, 2014.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the 2013 ACM annual conference on Human factors in computing systems*, pages 3267–3276. ACM.

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PloS one*, 6(12):e26752.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Inc. Google. 2014. Google flu trends. `http://www.google.org/flutrends`. Accessed on March 12, 2014.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers age and gender. In *Third International AAAI Conference on Weblogs and Social Media*.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

David A Huffaker and Sandra L Calvert. 2005. Gender, identity, and language use in teenage blogs. *Journal of Computer-Mediated Communication*, 10(2):00–00.

Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. I know what you did last summer: query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914. ACM.

Margaret L Kern, Johannes C Eichstaedt, H Andrew Schwartz, Gregory Park, Lyle H Ungar, David J Stillwell, Michal Kosinski, Lukasz Dziurzynski, and Martin EP Seligman. 2014. From sooo excited!!! to so proud: Using language to study development. *Developmental psychology*, 50(1):178–188.

Michal Kosinski and David J Stillwell. 2012. mypersonality project. *http://www.mypersonality.org/wiki/*.

Michal Kosinski, David J Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. volume 110, pages 5802–5805. National Acad Sciences.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.

Wendy Liu and Derek Ruths. 2013. Whats in a name? using first names as features for gender inference in twitter. In *Analyzing Microtext: 2013 AAAI Spring Symposium*.

Alessandra Marengoni, Sara Angleman, René Melis, Francesca Mangialasche, Anita Karp, Annika Garmen, Bettina Meinow, and Laura Fratiglioni. 2011. Aging with multimorbidity: a systematic review of the literature. *Ageing research reviews*, 10(4):430–439.

Robert R McCrae, Paul T Costa, Margarida Pedroso de Lima, António Simões, Fritz Ostendorf, Alois Angleitner, Iris Marušić, Denis Bratko, Gian Vittorio Caprara, Claudio Barbaranelli, et al. 1999. Age differences in personality across the adult life span: parallels in five cultures. *Developmental Psychology*, 35(2):466–477.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the demographics of twitter users. *ICWSM*, 11:5th.

Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. how old do you think i am?: A study of language and age in twitter. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *ICWSM*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *ICWSM*.

James W Pennebaker and Lori D Stone. 2003. Words of wisdom: language use over the life span. *Journal of Personality and Social Psychology*, 85(2):291–301.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. 71:2001.

2014. Social networking fact sheet. `http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/`. Accessed on August 26, 2014.

Francisco Rangel and Paolo Rosso. 2013. Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, page 177.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, volume 6, pages 199–205.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, Lyle Ungar, et al. 2013a. Characterizing geographic variation in well-being using tweets. In *ICWSM*.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David J Stillwell, Martin EP Seligman, et al. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Rong Su, James Rounds, and Patrick Ian Armstrong. 2009. Men and things, women and people: a meta-analysis of sex differences in interests. *Psychological Bulletin*, 135(6):859–884.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Benjamin Van Durme. 2012. Streaming analysis of discourse participants. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 48–58. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*.

Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, volume 13.