Controlling Human Perception of Basic User Traits

Daniel Preoţiuc-Pietro and **Sharath Chandra Guntuku** and **Lyle Ungar**
Positive Psychology Center
University of Pennsylvania
{danielpr@sas,sharathg@sas,ungar@cis}.upenn.edu

Abstract

Much of our online communication is text-mediated and, lately, more common with automated agents. Unlike interacting with humans, these agents currently do not tailor their language to the type of person they are communicating to. In this pilot study, we measure the extent to which human perception of basic user trait information – gender and age – is controllable through text. Using automatic models of gender and age prediction, we estimate which tweets posted by a user are more likely to mis-characterize his traits. We perform multiple controlled crowdsourcing experiments in which we show that we can reduce the human prediction accuracy of gender to almost random – an over 20% drop in accuracy. Our experiments show that it is practically feasible for multiple applications such as text generation, text summarization or machine translation to be tailored to specific traits and perceived as such.

1 Introduction

Advances in Natural Language Processing are leading to a point when text generation methods are deployed at scale. However, in the quest to make these applications more likable, effective and hence more usable, these methods should consider a way to adapt themselves to the person or type of persons they are interacting with (Bates, 1994; Loyall and Bates, 1997) e.g., a student may learn better from a tutoring agent that expresses similar traits to himself (Baylor and Kim, 2004).

In this study, we explore the feasibility of controlling human perception of traits using automated methods. Flekova et al. (2016); Carpenter et al. (2016) are the first to study the difference between user traits and their perception by external raters using tweets from social media. Their focus was on quantifying differences between perception and reality and analyzing text features which lead to mis-perception. This study goes a step further, and using the same experimental design and crowdsourcing, aims to use automatic methods to control human perception of basic user traits – here age and gender – through tweets. To this end, we use gender and age prediction algorithms to select tweets posted by users with a known trait with the goal of increasing or decreasing human rater accuracy in guessing their traits.

Obfuscating gender as identified by an automatic classifier was attempted in (Reddy and Knight, 2016). This problem is related, but very different to ours as we study human perception which is both different (Flekova et al., 2016) and more complex. Reddy and Knight (2016) study a range of lexical substitutions that can be performed in order to decrease the prediction accuracy of a classifier, although acknowledging that these may affect lexical coherence. In this pilot study, we circumvent this problem by using tweets known to have been written by the same person, with the downside of possible topic confounds.

Our experiments show that, for gender, we can decrease the human accuracy in perceiving gender from text by more than 20% as compared to a random selection of their tweets, with accuracy in this case being only slightly higher than chance. Further, this accuracy is even lower when predicting males. For age perception, we show consistent results in altering perception as both younger or older, albeit for relatively smaller age differences.

Applications of our proposed line of research include conversational agents or automated e-mail generation. Personalization was motivated in the context of machine translation (Mirkin et al.,

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2325–2331
Copenhagen, Denmark, September 7–11, 2017.

2015) and recently attempted for gender (Rabinovich et al., 2017), even though the authors do not use humans to evaluate perception of gender. Automatic text personalization to user traits can also go beyond basic demographics to more salient ones such as social status (Preoţiuc-Pietro et al., 2015a,b), political ideology (Preoţiuc-Pietro et al., 2017a) or psychological traits such as personality (Schwartz et al., 2013; Guntuku et al., 2015a,b, 2016, 2017), narcissism (Preotiuc-Pietro et al., 2016), trust or empathy (Abdul-Mageed et al., 2017).

2 Data Set

We study two user traits through two Twitter data sets containing users with known gender and age information. First, for gender, we use a subset of 200 users (100 males, 100 females) of the data set collected by (Burger et al., 2011) and released by (Volkova et al., 2013) which mapped users to their gender by linking their Twitter account to their publicly self-declared gender on related blogs. The age data set consists of 200 users that self-reported their age in a survey and disclosed their public Twitter data that are part of a larger set used in (Flekova et al., 2016). The users are chosen to have an age in the 15–34 year old interval in the year 2015 and we only use tweets posted in 2015 in our analysis. We selected exactly 10 users of each age in this interval, as these are the most frequent ages present in our data set, most language variation happens in this interval and these are the age range which raters can most accurately predict (Nguyen et al., 2014).

We use the Twitter API to download up to 3200 tweets from these users. We pre-process tweets by filtering those not written in English as detected by an automated method (Lui and Baldwin, 2012), removing duplicate tweets (i.e., having the same first 6 tokens) and removing re-tweets as these are not authored by the user. All potentially sensitive or revealing information contained in tweets such as URL's, usernames, @-mentions, phone numbers were removed and replaced with placeholders before shown to annotators. Other than publicly available tweets, no other metadata or information was presented with the task, so raters were not able to map the tweets to actual user identities. The raters were also unaware of the conditions (Random, Opposite, Same, Youngest or Oldest) they were assigned to when performing the ratings. All our experiments received approval from Institutional Review Board (IRB) of the University of Pennsylvania.

We are aware that the proposed long-term applications we envision for this research can have personal impact on users. Hence, we propose following criteria which should be at the core of future research in controlling human perception, which we encourage to be completed over time:

- **Transparency**: data trained to build the personalized models should be transparent to any user. This would allow to observe any possible biases that may exist in the data.
- **Control**: the user interacting with a personalized system should be aware of the type of personalization employed by the agent (e.g. by gender, by which particular age group) and should be able to disable it when desired.

3 Experimental Setup

We use Amazon Mechanical Turk to create crowdsourcing tasks for predicting age and gender from tweets. Each HIT consists of 20 tweets authored by a single user and selected using different methods. The annotators were asked to predict gender (M/F) or age (integer value in 13–90) and rate their confidence of their guess from 1 (not at all confident) to 5 (very confident). We collected 3 annotations for each author and set of tweets.

Participants received a small compensation (.04$) for each rating and could repeat the task as many times as they wished, but never for the same authors and set of tweets. They were also presented with an initial bonus (.25$). For quality control, the participants underwent a short training and qualification questions, their location was limited to the US and they had to spend at least 10 seconds on each HIT before they were allowed to submit their guess.

In order to estimate which tweets are more likely to be written by females or a older user, we use the classifier introduced in (Sap et al., 2014). This is a regularized Linear SVM that obtains state-of-the-art prediction results on user gender (91.9% accuracy) and age ($r = .835$) prediction from social media text. We apply the model to all our tweets and select for each user 20 tweets based on the following criteria:

- **Random**: tweets chosen at random from a user's timeline;
- **Opposite**: for gender, tweets that are predicted

	Single Rating	Majority Vote
Baseline	50%	50%
Opposite	55.74%	61.57%
Random	76.67%	83.99%
Same	91.33%	95.49%

Table 1: Human accuracy in gender perception experiments in the three text selection conditions.

as more likely to be written by someone of the different gender;

- **Same**: for gender, tweets predicted to be written by someone of the same gender as the author.
- **Youngest**: for age, tweets from a user that are predicted as youngest age;
- **Oldest**: for age, tweets from a user that are predicted as oldest age;

The tweets selected based on the automatic prediction are presented in the order of prediction scores e.g. tweets for **Youngest** are sorted with the lowest predicted age being shown at the top of the list. Experiments with random ordering of tweets showed similar results.

4 Results

In this section we analyze the extent to which our experiments manage to alter trait perception, the errors and confidence of the annotations.

4.1 Gender

Overall accuracy results for our gender experiments are presented for both individual ratings and majority vote in Table 1. In all experimental setups, the raters were able to guess gender better than chance, with the majority vote of the three raters higher by a significant margin (5.77% on average) than the individual votes.

Our selection procedure has great impact on rater accuracy. Selecting tweets most likely to be written by the opposite gender – even if they are posted by the same user in reality – has an impact of decreasing the individual rater accuracy by 20.93% to only slightly above random guess (55.75%). For the majority vote ratings, the decrease is 22.42% (paired T-test, $t = 8.06$, $p < 10^{-14}$). On the other hand, selecting the tweets that are most likely to be posted by a user from the same gender as determined by our automatic model has the impact of increasing the individual rater accuracy by 14.66%. The majority vote prediction is increased by a relatively smaller amount (11.5% – paired T-test, $t = 7.09$, $p < 10^{-11}$), which we attribute to the accuracy being very close to oracle performance.

The confusion matrices from the three experiments are presented in Table 2. A couple of patterns stand out: females are easier to be accurately identified in all three experiments and males are more likely to be confused for females than vice-versa. This resulted in raters guessing more users to be females, despite our data set being balanced. Intriguingly, in the **Opposite** experiment, males were more often confused for females than correctly guessed, with females being guessed far more accurately, making the average accuracy better than chance. In the **Same** experiment, females are again easier to guess, with accuracy being very close to perfect. These results show that females are more distinctive in their language use on Twitter and thus are harder to be confused for males. On the other hand, as proven by the **Opposite** experiment, posts written by males can be selected such that they are perceived as written by females.

The inter-annotator agreement is presented in Table 4. Pairwise agreement at a user level is very high for the **Same** setup, decreasing significantly for the **Random** and **Opposite** setups.

The average self-rated confidence in assessments for the three experiments are presented in Table 3. Self-rated confidence mirrors almost perfectly the accuracy scores in all experiments and cases: confidence is higher on average in cases when accuracy is higher. Users are in general more confident when accurately guessing a female, and are least accurate when inaccurately guessing a female. Noteworthy, in the **Opposite** experiment, users who incorrectly guessed males were more confident than when correctly identifying males, which is not the case for females. This further shows that females are use more distinctive language on Twitter, while males could be more easily mistaken for females.

4.2 Age

Overall accuracy results for our age experiments are presented in Table 5. We only report results with a user age computed as the average three guesses. Results with individual ratings are very similar and we omit them for brevity.

The experiments show that our model's selection matches human perception: in the **Younger** experiment, the average predicted age is lower than in the **Random** experiment, which is in turn lower on average than the predicted age in the **Oldest** setup. Further, in the **Younger** experiment, many

(a) Random tweets (Acc. 76.66%).

Real	Predicted: Male	Predicted: Female
Male	35.99%	14%
Female	9.33%	40.67%

(b) Opposite Tweets (Acc. 55.73%).

Real	Predicted: Male	Predicted: Female
Male	23.47%	26.35%
Female	17.9%	32.26%

(c) Same Tweets (Accuracy 91.33%).

Real	Predicted: Male	Predicted: Female
Male	43.5%	6.5%
Female	2.16%	47.83%

Table 2: Normalized confusion matrices of human guesses (**Predicted**) compared to ground truth (**Real**).

(a) Random tweets (Average 3.37).

Real	Predicted: Male	Predicted: Female
Male	3.22	2.92
Female	2.78	3.80

(b) Opposite Tweets (Average 3.44).

Real	Predicted: Male	Predicted: Female
Male	3.24	3.57
Female	3.14	3.65

(c) Same Tweets (Average 3.85).

Real	Predicted: Male	Predicted: Female
Male	3.70	2.76
Female	2.76	4.18

Table 3: Average confidence of human guesses (**Predicted**) depending on ground truth (**Real**).

(a) Gender

	Cohen's κ	Agreement
Opposite	.252	64.3%
Random	.354	68%
Same	.764	88.3%

(b) Age

	St. Dev.	Pearson r
Youngest	3.40	.368
Random	4.06	.170
Oldest	4.89	.341

Table 4: Inter-annotator agreement statistics.

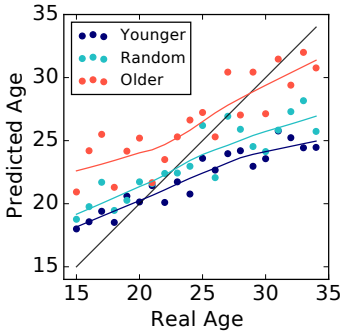

Figure 1: The average predicted ages compared to real age in the three experiments. The black line represents the ideal fit, the colored lines represent a LOESS fit to the data.

more users' age is under-estimated as compared to when predicting average age and in the **Older** experiment, more users' age is over-estimated. We also note that in the **Random** setup, raters tend to under-estimate age (53.5% younger vs. 39.3% older), with the mean being lower than in the data (23.3 vs. 24.5), which aligns with previous research (Nguyen et al., 2014).

Figure 1 plots the average prediction for users by age in the three experiments. Intriguingly, even in the **Younger** setup, users under 18 y.o. are predicted as older, while the groups of users over 20 y.o. are all under-predicted. Notably, the same near-linear pattern largely holds for the other two experiments, with the age cut-off being different (23 for **Random**, 27 for **Oldest**).

The accuracies of the three experiments are very similar regardless if comparing the number of correct guesses or guesses within 1, 3 of 5 years of the actual age. By examining Figure 1, we realize that the set of users accurately predicted shifts from one method to the other. This highlights that, even if controlling age perception is feasible, this is possible only for a difference of a few years.

The inter-annotator agreement is presented in Table 4. First, the average standard deviation across the three guesses for each author shows that **Youngest** setup generates the most similar guesses, which tend to be in the younger age range. In contrast, the **Oldest** setup generates the largest variance in guesses. Average Pearson correlation between the three guesses per author shows that both controlled setups result in higher agreement between raters than the **Random** setup, which shows that users are easier to rank by age based on their extreme language use (**Oldest** or **Youngest**) compared to a random tweet sample.

Finally, the average self-confidence of the ratings is highest in the **Youngest** experiment ($\mu = 3.35$), followed by the **Older** experiment ($\mu = 3.20$) with the **Random** experiment ($\mu = 2.97$) lowest. Further, we checked if there is a relationship between true or predicted age and self-rated confidence. In the **Youngest** experiment, both true age and predicted age are negatively correlated with self-rated confidence (true age: Pearson $r = -0.218$, p-value $< 10^{-8}$, predicted age: Pearson $r = -.246$, p-value $< 10^{-10}$), showing that raters believe their guess is easier when encountering younger users. In the **Random** experiment, a significant correlation exists between self-rated confi-

	Mean	σ	Median	Correct	Younger	Older	$\leq$1 Year	$\leq$3 Years	$\leq$5 Years
Baseline	24.0	0.0	24.0	5.0%	40.0%	45.0%	15.0%	35.0%	55.0%
Youngest	22.0	4.4	21.6	7.6%	62.2%	30.1%	20.7%	44.3%	59.0%
Random	23.3	4.8	22.8	7.1%	53.5%	39.3%	21.3%	42.8%	61.3%
Oldest	26.4	5.7	26.0	7.5%	36.2%	56.2%	22.0%	41.5%	60.6%

Table 5: Age prediction results in the three experimental setups. The predicted user age is the average age of the three human ratings. The **Baseline** represents always selecting the average age in our data set.

Gender			
Opposite (M)		**Opposite (F)**	
dress	Just saw a **dress** style i can only describe as [...]	**his**	Chinese men amputated **his** own leg: URL
herself	What's USER doing on The Voice and why is she calling **herself** 'James'	**wife**	The Good **Wife** exists to squeeze every last ounce of sincerity and hope out of your soul.
husband	.USER USER Wait ... you're meeting your **husband** in Cleveland?	**haircut**	Right. **Haircut** when I get home.
women's	Forget **women's** rights or voters rights [...]	**burger**	Best **burger** anywhere. [...]
him	Glad I was able to send **him** to the heartbreak hotel...	**of**	I spend 99% **of** my awake time thinking about food I can't eat.
Age			
Youngest (–)		**Oldest (+)**	
literally	that's **literally** living the dream	**daughter**	[...] USER makes my 2-month-old **daughter** stop fussing :)
so	im laughing **so** hard	**years**	i love [...] regretting everything from like three **years** ago
though	You cute **though**	**via**	Christmas is almost here! Let's party! URL **via** USER
excited	IM SO **EXCITED** URL	**ago**	One year **ago** today URL URL
guys	you **guys** killed it USER	**ok**	I'm **OK** with that. URL

Table 6: Most impactful features in tweet selection and representative tweet.

dence and predicted age only (Pearson $r = -.172$, p-value $< 10^{-5}$), while we found no relationship in the **Oldest** experiment. This indicates that language use at least apparently is more distinguishable for younger users, probably due to specific topics or interests.

5 Qualitative Analysis

Finally, we show in Table 6 the top features that impact selection of tweets in representative setups from this paper together with a representative tweet. The top features are computed by multiplying the regression/classification weight with the user-normalized average frequency of the feature in the displayed tweets. For gender, we use the **Opposite** setup to show words most indicative of females present in tweets selected and written by males and viceversa. Gender specific features are used with different senses than usual ('dress', 'wife', 'women's'), in reference to other persons rather than oneself ('herself', 'his'), or represent stylistic ('of') or topical ('haircut', 'burger') differences. For age, we select the feature most indicative of a younger user in the **Youngest** setup and the ones most indicative of older age in the **Oldest** setup. In this case, most of the top words are stylistic ('literally', 'so', 'though', 'excited, 'guys', 'ok', 'via') with features indicative of older age referencing the past ('years', 'ago') or generally specific of older age ('daughter').

6 Conclusions

We have presented the first study into automatically controlling human perception of written text. Our exploration used gender and age as basic human traits, which most have a good level of knowledge about, to measure the extent to which altering perception in text-mediated communication is feasible. Our results showed that this is possible to some extent, being especially accurate for males. Age experiments demonstrated consistent results across the three experiments, although alteration seems possible only for relatively small age deltas.

In this first experiments on this topic, we chose to perform tweet selection rather than generation, as these methods often generate text that is not semantically and syntactically correct or natural for a reader. In future work, we will experiment with automatically altering or generating text while keeping topic constant, as our current results are in part topically driven. Alterations can be performed through stylistic transformations such as normalization or by using paraphrasing as suggested in (Preoţiuc-Pietro et al., 2016, 2017b).

Text adaptation is especially important for conversational agents that interact only through text. As humans, we automatically perform this adaptation through multiple additional channels: speech tone, frequency, facial expression; which the agent can not alter. In addition to methodology, future work will also need to take into account ethical implications of this personalization.

Acknowledgments

This project/publication was made possible through the support of a grant from Templeton Religion Trust (TRT-0048). The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Templeton Religion Trust.

References

Muhammad M. Abdul-Mageed, Anneke Buffone, Hao Peng, Giorgi Salvatore, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing Pathogenic Empathy in Social Media. In *Proceedings of the Eleventh International AAAI Conference on the Web and Social Media*, ICWSM, pages 448–451.

Joseph Bates. 1994. The Role of Emotion in Believable Agents. *Communications of the ACM*, 37(7):122–125.

Amy L Baylor and Yanghee Kim. 2004. Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role. *Intelligent Tutoring Systems. ITS 2004. Lecture Notes in Computer Science*, 3220:592–603.

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1301–1309.

Jordan Carpenter, Daniel Preoţiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret Kern, Anneke Buffone, Lyle Ungar, and Martin Seligman. 2016. Real Men don't say 'cute': Using Automatic Language Analysis to Isolate Inaccurate Aspects of Stereotypes. *Social Psychological and Personality Science*, 8:310–322.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing Biases in Human Perception of User Age and Gender from Text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 843–854.

Sharath Chandra Guntuku, Weisi Lin, Jordan Carpenter, Wee Keong Ng, Lyle H Ungar, and Daniel Preoţiuc-Pietro. 2017. Studying personality through the content of posted and liked images on twitter. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 223–227. ACM.

Sharath Chandra Guntuku, Weisi Lin, Michael James Scott, and Gheorghita Ghinea. 2015a. Modelling the influence of personality and culture on affect and enjoyment in multimedia. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 236–242. IEEE.

Sharath Chandra Guntuku, Lin Qiu, Sujoy Roy, Weisi Lin, and Vinit Jakhetiya. 2015b. Do others Perceive you as you want them to?: Modeling Personality based on Selfies. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*, pages 21–26. ACM MM.

Sharath Chandra Guntuku, Joey T Zhou, Sujoy Roy, Lin Weisi, and Ivor W Tsang. 2016. Who likes What, and Why? Insights into Personality Modeling based on Image 'Likes'. *IEEE Transactions on Affective Computing*, PP:1–14.

A Bryan Loyall and Joseph Bates. 1997. Personality-rich Believable Agents that use Language. In *Proceedings of the First International Conference on Autonomous Agents*, AGENTS, pages 106–113.

Marco Lui and Timothy Baldwin. 2012. Langid.Py: An Off-the-shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, ACL, pages 25–30.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating Personality-aware Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1102–1108.

Dong-Phuong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING, pages 1950–1961.

Daniel Preoţiuc-Pietro, Jordan Carpenter, and Lyle Ungar. 2017a. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL.

Daniel Preoţiuc-Pietro, Jordan Carpenter, and Lyle Ungar. 2017b. Personality Driven Differences in Paraphrase Preference. In *Proceedings of the Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, ACL.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics*, ACL, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*, 10(9).

Daniel Preoţiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering User Attribute Stylistic Differences via Paraphrasing. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI, pages 3030–3037.

Daniel Preotiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the Dark Triad of Personality using Twitter Behavior. In *Proceedings of the 25th ACM Conference on Information and Knowledge Management*, CIKM.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2017. Personalized Machine Translation: Preserving Original Author Traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 1074–1084.

Sravana Reddy and Kevin Knight. 2016. Obfuscating Gender in Social Media Writing. In *Workshop on Natural Language Processing and Computational Social Science (NLP for CSS)*, EMNLP, pages 17–26.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and H Andrew Schwartz. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PLoS One*, 8.

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 in Natural Language Processing*, EMNLP, pages 1815–1827.