Analyzing Biases in Human Perception of User Age and Gender from Text

Lucie Flekova[*]
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Jordan Carpenter and **Salvatore Giorgi**
Positive Psychology Center
University of Pennsylvania

Lyle Ungar and **Daniel Preoţiuc-Pietro**
Computer & Information Science
University of Pennsylvania

Abstract

User traits disclosed through written text, such as age and gender, can be used to personalize applications such as recommender systems or conversational agents. However, human perception of these traits is not perfectly aligned with reality. In this paper, we conduct a large-scale crowdsourcing experiment on guessing age and gender from tweets. We systematically analyze the quality and possible biases of these predictions. We identify the textual cues which lead to miss-assessments of traits or make annotators more or less confident in their choice. Our study demonstrates that differences between real and perceived traits are noteworthy and elucidates inaccurately used stereotypes in human perception.

1 Introduction

There are notable differences between actual user traits and their perception by others (John and Robins, 1994; Kobrynowicz and Branscombe, 1997). Assessments of the perceived traits are dependent, for example, on the interpretation skills of a judge (Kenny and Albright, 1987) and the ability of users to deliberately adjust their behavior to the way they intend to be perceived e.g., for following a social goal (Kanellakos, 2002). People typically use stereotypes – a set of beliefs, generalizations, and associations about a social group – to make judgements about others. The discrepancy between stereotypes and actual group differences is an important topic in psychological research (Eagly, 1995; Dovidio et al., 1996; John and Robins, 1994; Kobrynowicz and Branscombe, 1997). Such differences are likely reflected through one's writing.

With the Internet a substantial part of daily life, users leave enough footprints which allow algorithms to learn a range of individual traits, some with even higher accuracy than the users' own family (Youyou et al., 2015). With an increase in readily available user generated content, prediction of user attributes has become more popular than ever. Researchers built learning models to infer different user traits from text, such as age (Rao et al., 2010), gender (Burger et al., 2011; Flekova and Gurevych, 2013), location (Eisenstein et al., 2010), political orientation (Volkova et al., 2014), income (Preoţiuc-Pietro et al., 2015c), socio-economic status (Lampos et al., 2016), popularity (Lampos et al., 2014), personality (Schwartz et al., 2013) or mental illnesses (De Choudhury et al., 2013; Coppersmith et al., 2014; Preoţiuc-Pietro et al., 2015a).

Prediction models are trained on large data sets with labels extracted either from user self-reports (Preoţiuc-Pietro et al., 2015b) or perceived from annotations (Volkova et al., 2015; Volkova and Bachrach, 2015). The former is useful in obtaining accurate prediction models for unknown users while the latter is more suitable in applications that interact with humans. Previous studies showed the implications of perceived individual traits to the believability and likability of autonomous agents (Bates, 1994; Loyall and Bates, 1997; Baylor and Kim, 2004).

This study aims to emphasize the differences between real user traits and how these are perceived by humans from Twitter posts. In this context, we address the following research questions:

[*] Project carried out during a research stay at the University of Pennsylvania

- How accurate are people at judging traits of other users?
- Are there systematic biases humans are subject to?
- What are the implications of using human perception as a proxy for truth?
- Which textual cues lead to a false perception of the truth?
- Which textual cues make people more or less confident in their ratings?

We use age and gender as target traits for our analysis, as these are considered basic categories in person assessment (Quinn and Macrae, 2005) and are highly studied by previous research. Using a large-scale crowdsourcing experiment, we demonstrate that human annotators are generally accurate in assessing the traits of others. However, they make systematically different types of errors compared to a prediction model trained using the bag-of-words assumption. This hints at the fact that annotators over-emphasize some linguistic features based on their stereotypes. We show how this phenomenon can be leveraged to improve prediction performance and demonstrate that by replacing self-reports with perceived annotations we introduce systematic biases into our models.

In our analysis section, we directly test the accuracy of these stereotypes, as the human predictions must rely on these theories of relative differences between groups if no explicit cues are mentioned. We uncover remarkable differences between actual and perceived traits by using multiple lexical features: unigrams, clusters of words built from word embeddings and emotions expressed through posts. In our analysis of features that lead to wrong assessments we uncover that humans mostly rely on accurate stereotypes from textual cues, but sometimes over-emphasize them. For example, annotators assume that males post more than they do about sports and business, females show more joy, older users more interest in politics and younger users use more slang and are more self-referential. Similarly, we highlight the textual features which lead to higher self-reported confidence in guesses, such as the mentions of family and beauty products for gender or college and school related topics for age.

2 Related Work

Studying gender differences has been a popular psychological interest over the past decades (Gleser et al., 1959; McMillan et al., 1977). Traditional studies worked on small data sets, which sometimes led to contradictory results – (Mulac et al., 1990) cf. (Pennebaker et al., 2003). Over the past years, researchers discovered a wide range of gender differences using large collections of data from social media or books combined with more sophisticated techniques. For example, Schler et al. (2006) apply machine learning techniques to a corpus of 37,478 blogs from the Blogger platform and find differences in the topics males and females discuss. Newman et al. (2008) showed that female authors are more likely to include pronouns, verbs, references to home, family, friends and to various emotions. Male authors use longer words, more articles, prepositions and numbers. Topical differences include males writing more about current concerns (e.g., money, leisure or sports). More recent author profiling experiments (Rangel et al., 2014; Rangel et al., 2015) revealed that gender can be well predicted from a large spectrum of textual features, ranging from paraphrase choice (Preoţiuc-Pietro et al., 2016), emotions (Volkova and Bachrach, 2016), part-of-speech (Johannsen et al., 2015) and abbreviation usage to social network metadata, web traffic (Culotta et al., 2015), apps installed (Seneviratne et al., 2015) or Facebook likes (Kosinski et al., 2013). Bamman et al. (2014) also examine individuals whose language does not match their automatically predicted gender. Most of these experiments were based on self-reported gender in social media profiles.

The relationship between age and language has also been extensively studied by both psychologists and computational linguists. Schler et al. (2006) automatically classified blogposts into three age groups based on self-reported age using features from the Linguistic Inquiry and Word Count Framework (Pennebaker et al., 2001), online slang and part-of-speech information. Rosenthal and McKeown (2011) analyzed how both stylistic and lexical cues relate to gender on blogs. On Twitter, Nguyen et al. (2013) analyzed the relationship between language use and age, modelled as a continuous variable. They found similar language usage trends for both genders, with increasing word and tweet length with age, and an increasing tendency to write more grammatically correct, standardized text. Flekova et al. (2016) identified age specific differences in writing style and analyzed their impact beyond income. Recently, Nguyen et al. (2014) showed that age prediction is more difficult as age increases, specifically over 30 years. Hovy and Søgaard (2015) showed that the author age is a factor influencing training part-of-speech taggers.

Recent results on social media data report a performance of over 90% for gender classification and a correlation of $r \sim 0.85$ for age prediction (Sap et al., 2014). However, authors can introduce their biases in text (Recasens et al., 2013). Accurate prediction of the true user traits is important for applications such as recommender systems (Braunhofer et al., 2015) or medical diagnoses (Chattopadhyay et al., 2011). Influencing perceived traits, on the other hand, enables a whole different range of applications - for example, researchers demonstrated that the perceived demographics influence student attitude towards a tutor (Baylor and Kim, 2004; Rosenberg-Kima et al., 2008). Perception alterations do not only strive for likeability - people intentionally use linguistic nuances to express social power (Kanellakos, 2002), which can be recognized by computational means (Bramsen et al., 2011). McConnell and Fazio (1996) show how gender-marked language colors the perception of target personality characteristics – enhanced accessibility of masculine and feminine attributes brought about by frequent exposure to occupation title suffixes influences the inferences drawn about the target person.

3 Data

In this study, we focus on analyzing human perception of two user traits: gender and age. For judging, we build data sets using publicly available Twitter posts from users with known self-reported age and gender. To study gender, we use the users from Burger et al. (2011), which are mapped to their self-identified gender as mentioned in other user public profiles linked to their Twitter account. This data set consists of 67,337 users, from which we subsample 2,607 users for human assessment. The age data set consists of 826 users that self-reported their year of birth and Twitter handle as part of an online survey.

We use the Twitter API to download up to 3200 tweets from these users. These are filtered for English language using an automatic method (Lui and Baldwin, 2012) and duplicate tweets are eliminated (i.e., having the same first 6 tokens) as these are usually generated automatically by apps. Tweet URLs and @-mentions are anonymized as they may contain sensitive information or cues external to language use. For human assessment, we randomly select 100 tweets posted in the same 6 month time interval from the users where gender is known. For the users of known age we randomly select 100 tweets posted during the year 2015.

4 Experimental Setup

We use Amazon Mechanical Turk to create crowdsourcing tasks for predicting age and gender from tweets. Each HIT consists of 20 tweets randomly sampled from the pool of 100 tweets of a single user. Each user was assessed independently by 9 different annotators. Using only these tweets as cues, the annotators were asked to predict either age (integer value) or gender (forced choice binary male/female) and self-rate the confidence of their guess on a scale from 1 (not at all confident) to 5 (very confident).

Participants received a small compensation (.02$) for each rating and could repeat the task as many times as they wished, but never for the same author. They were also presented with an initial bonus (.25$) and a similar one upon completing a number of guesses. For quality control, we used a set of HITs where the user's age or gender was explicitly stated within the top 10 tweets displayed in the task. The control HIT appeared 10% of the time and all annotators missing the correct answer twice were excluded from annotation and all their HITs invalidated. A total of 28 annotators were banned from the study. Further, we limited annotator location to the US and they had to spend at least 10 seconds on each HIT before they were allowed to submit their guess.

5 Crowdsourcing Results

We first analyze the annotator performance on the gender and age prediction tasks from text. For gender, individual ratings have an overall accuracy of 75.7% (78.3% for females and 72.8% for males). The pairwise inter-annotator agreement for 9 annotators is 70.0%, Fleiss' Kappa 39.6% and Krippendorf's Alpha 39.6%, while keeping in mind that the annotators are not the same for all Twitter users. In terms of confidence, average self-rated confidence for correct guesses is $\mu = 3.47$, while average confidence for wrong guesses is $\mu = 2.84$. In total, 1083 individual annotators performed an average of $\mu = 22.3$ ratings with the standard deviation $\sigma = 32.76$ and the median of 12.

We use the majority vote as the method of label aggregation for gender prediction. The majority vote accuracy on predicting the gender of Twitter users is 85.8% with the majority class baseline being 51.9% female, a result comparable to a previous study (Nguyen et al., 2014). Table 1a presents the gender confusion matrix. Female users were more often classified into a correct class (88.3% recall for females cf. 83.5% for males). The majority of errors was caused by male users mislabeled as female. This results in higher precision on classifying male users (86.9% cf. 85.3% for females). In terms of overall self-reported confidence of the annotators, decisions on actual female users were on average more confidently rated ($\mu = 3.60$) compared to males ($\mu = 3.31$), which is in consensus with higher accuracy for females. Figure 2 shows the relationship between annotation accuracy and average confidence per Twitter users. The relationship is non-linear, with the average confidence in the 1–3 range for gender having little impact on the prediction accuracy.

For the age annotations, the correlation between predicted and real age for individual ratings is $r = 0.416$. The mean absolute error (MAE) is 7.31, while the baseline MAE obtained if predicted the sample mean real age is 8.61. The intraclass correlation coefficient between the 9 ratings is 0.367 and taking into account the fact that the annotators were different across users (Shrout and Fleiss, 1979), while the average standard deviation of the 9 user guesses for a single Twitter user is $\sigma = 5.60$. Individual rating confidence and the Mean Absolute Error (MAE) are anti-correlated with $r = -0.112$, matching the expectation that higher self-reported confidence leads to lower errors. The 691 different annotators performed on average $\mu = 10.68$ ratings with standard deviation $\sigma = 21.95$ and a median of only 4 ratings. Based on feedback, this was due to the difficulty of the age task.

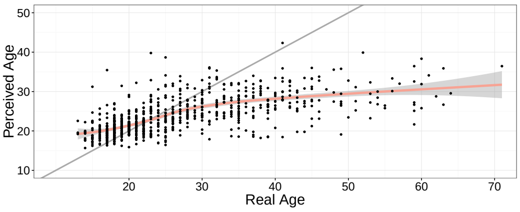

Figure 1: Real age predictions compared to average predicted age. The line shows a LOESS fit.

In the rest of the age experiments, we consider the predicted age of a user as a mean of the 9 human guesses. Overall, the correlation between average predicted age and real age is $r = 0.631$. The MAE of the average predicted age is 6.05. MAE and average self-rated confidence by user are negatively correlated with $r = -0.21$. Figure 3 plots annotation confidence on a Twitter user level and MAE of age guesses. Again, the relationship between confidence and MAE is non-linear, with confidences of 1–2 having similar average MAE, with the error decreasing as the average of the confidence ratings per author is higher. Figure 1 shows a scatter plot comparing real and predicted age together with a non-linear fit of the data. From this figure, we observe that annotators under-predict age, especially for older users. The correlation of MAE with real age is very high ($r = 0.824$) and the residuals are not normally distributed.

Figures 4 and 5 show the accuracy if only a subsample of the ratings is used and the labels are aggregated using majority vote for gender and using average ratings for age. For gender, we notice that accuracy abruptly increases from 1 to 3 votes and to a lesser extent from 3 to 5 votes, but the differences between 5, 7 and 9 votes are very small. Similarly, for age, MAE decreases up until using 4 guesses, where it reaches a plateau. These experiments suggest that a human perception accuracy can be sufficiently approximated using up to 5 ratings - additional annotations after this point have negligible contribution.

Finally, the individual annotator accuracy is independent on the number of users rated. For gender, the Pearson correlation between accuracy and number of ratings performed is $r = .009$ ($p = .75$) and for age the Pearson correlation between MAE and the number of ratings performed by a user is $r = -.013$ ($p = .71$). This holds even when excluding users who performed few ratings.

6 Uncovering Systematic Biases

In this section, we use the extended gender data set in order to investigate if human guesses contain systematic biases by comparing these guesses to those from a bag-of-words prediction model. We then test what is the impact of using human guesses as labels and if human ratings offer additional in-

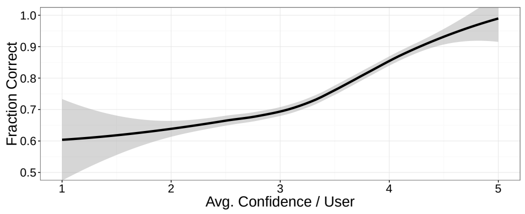

Figure 2: Gender – Fraction of correct guesses as a function of average confidence per rated Twitter user. Black line shows a LOESS fit.

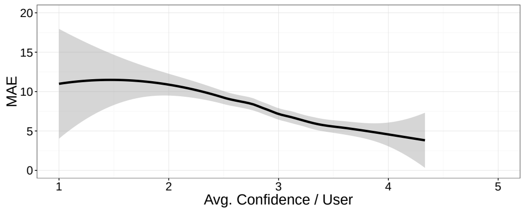

Figure 3: Age – Mean Absolute Error as a function of average confidence per rated Twitter user. Black line shows a LOESS fit.

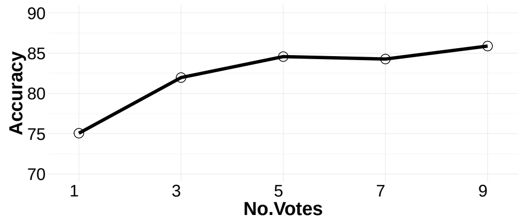

Figure 4: Gender – Majority vote accuracy based on number of annotator guesses aggregated.

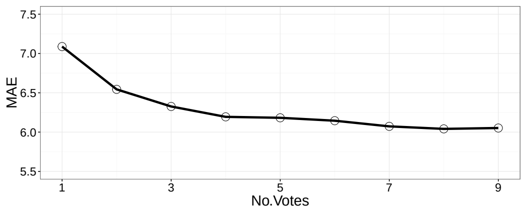

Figure 5: Age – Average Mean Absolute Error based on number of annotator guesses aggregated.

formation to predictive models.[1]

6.1 Comparison to Bag-of-Words Predictions

First, we test the hypothesis that annotators emphasize certain stereotypical words to make their guesses. To study their impact, we compare human guesses with those from a statistical model using the bag-of-words assumption for systematic differences. The automatic prediction method using bag-of-words text features offers a generalisation of individual word usage patterns shielded from biases.

We use Support Vector Machines (SVM) with a linear kernel and ℓ_1 regularization (Tibshirani, 1996), similarly to the state-of-the-art method in predicting user age and gender (Sap et al., 2014). The features for these models are unigram frequency distributions computed over the aggregate set of messages from each user. Due to the sparse and large vocabulary of social media data, we limit the unigrams to those used by at least 1% of users.

We train a classifier on a balanced set of 11,196 Twitter users from our extended data set. We test on the 2,607 users rated by the annotators using only the 100 tweets the humans had access when making their predictions. Table 1b shows the system performance reaching an accuracy of 82.9%, with the human performance on the same data at 85.88%. In contrast to the human prediction, the precision is higher for classifying females (84.9% cf. 80.9% for males) and the recall is higher for males (85.4% cf. 80.4% for female). This is caused by both higher classifier accuracy for males and by a switch in rank between the type I and type II errors.

In Table 1c we directly compare the human and automatic predictions, highlighting that 13.6% of the labels are different. Moreover, there is an asymmetry between the tendency of humans to mislabel males with females and the classifier. This leads to the conclusion that humans are sensitive to biases which we will qualitatively investigate in the following sections.

6.2 Human Predictions as Labels

Previously, we have shown that perceived annotated traits are different in many aspects to actual traits. To quantify their impact, we use these labels for training two classifiers and compare them on predicting the true gender for unseen users.

Both systems are trained on the 260,700 messages from 2,607 users and only differ in the labels assigned to users: majority annotator vote or self-reports. Results on the held-out set of 11,196 users (of which 6,851 males and 7,596 females) are presented in Table 2. The system trained on real labels outperforms that trained on perceived ones (accuracy of 85.32% cf. 83.40%). Furthermore, in the system trained on perceived labels, the same type of error as for the human annotation is more prevalent and is overemphasized compared to our

[1] Experiments for age could not be replicated due to insufficient labeled users.

(a) Majority annotator vote.

		Pred.H	
		Male	**Female**
Real	**Male**	40.1%	7.9%
	Female	6.1%	45.8%

(b) Classifier.

		Pred.C	
		Male	**Female**
Real	**Male**	42.2%	7.2%
	Female	9.9%	40.7%

(c) Classifier compared to majority annotator vote.

		Pred.H	
		Male	**Female**
Pred.C	**Male**	40.3%	8.0%
	Female	5.6%	46.1%

Table 1: Normalized confusion matrices of human annotations (**Pred.H**) to ground truth (**Real**), classifier performance (**Pred.C**) to ground truth (**Real**), and human annotations (**Pred.H**) to classifier performance (**Pred.C**) on the same data set.

previous results – males are predicted with high precision (85%) but low recall (79%) and many of them are misclassified as women. In the system trained on ground truth, both types of errors are more balanced with more males classified correctly – similar precision (84%) but higher recall (86%).

6.3 Combining Human and Automatic Predictions

We have shown that human perceived labels and automatic methods capture different information. This information may be leveraged to obtain better overall predicting performance. We test this by using a linear model that combines two features: the human guesses – measured as the proportion of guesses for female – and classifier prediction – binary value. Even this simple method of label combination obtains a classification accuracy of 87.7%, significantly above majority vote of human guesses (85.8%) and automatic prediction (82.9%) individually. This demonstrates that both methods can complement each other if an increase in accuracy is needed.

(a) Trained on perceived gender. Accuracy = 83.4%

		Pred.	
		Male	**Female**
Real	**Male**	37.5%	9.9%
	Female	6.6%	45.9%

(b) Trained on actual gender. Accuracy = 85.3%

		Pred.	
		Male	**Female**
Real	**Male**	40.5%	6.9%
	Female	7.8%	44.7%

Table 2: Normalized confusion matrices for system comparison when using perceived or ground truth labels.

7 Textual Differences between Perceived and Actual Traits

We have so far demonstrated that differences exist between the human perception of traits and real traits. Further, human errors differ systematically from a statistical model which generalizes word occurrence patterns. In this section, we directly identify the textual cues that bias humans and cause them to mislabel users.

In addition to unigram analysis, in order to aid interpretability of the feature analysis, we group words into clusters of semantically similar words or *topics* using a method from (Preoţiuc-Pietro et al., 2015b). We first obtain word representations using the popular skip-gram model with negative sampling introduced by Mikolov et al. (2013) and implemented in the Gensim package (layer size 50, context window 5). We train this model on a **separate** reference corpus containing $\sim$ 400 million tweets. After computing the word vectors, we create a word × word semantic similarity matrix using cosine similarity between the vectors and group the words into clusters using spectral clustering (Shi and Malik, 2000). Each word is only assigned to one cluster. We choose a number of 1,000 topics based on preliminary experiments. Further, we use the NRC Emotion Lexicon (Mohammad and Turney, 2013) to measure eight *emotions* (anger, fear, anticipation, trust, surprise, sadness, joy and disgust) and two *sentiments* (negative and positive). A user's score in each of these 10 dimensions is represented as a weighted sum of its words multiplied by their lexicon score.

7.1 Gender Perception

To study gender perception, we first define a measure of *perceived gender expression*, calculated as the fraction of female guesses out of the 9 guesses for each Twitter user. We then compute univariate correlations the text-derived features and the user

Perceived – Female				**Perceived – Male**			
Topic	**Perc**	**Real**	**Cont**	Topic	**Perc**	**Real**	**Cont**
such, loving, pretty, beautiful, gorgeous	.416	.348	.176	nation, held, rally, defend, supporters	-.372	-.281	-.176
bed, couch, blanket, lying, cozy	.424	.376	.165	players, teams, crowds, athletes, clubs	-.370	-.284	-.171
hair, blonde, shave, eyebrows, dye	.379	.325	.152	training, team, field, coach, career	-.323	-.246	-.148
friend, boyfriend, bf, bff, gf	.365	.308	.149	heat, game, nba, lakers, playoff	-.314	-.237	-.145
girl, lucky, she's, you're, he's	.378	.336	.143	draft, trade, deadline, stat, retire	-.303	-.223	-.143
sweet, angel, honey, pumpkin, bunny	.365	.322	.138	ref, offensive, foul, defensive, refs	-.324	-.255	-.142
cleaning, laundry, packing, dishes, washing	.350	.307	.133	second, third, grade, century, period	-.282	-.195	-.142
awake, dream, sleep, asleep, nights	.327	.276	.130	former, leader, chief, vice, minister	-.316	-.244	-.142
cry, heart, smile, deep, whenever	.331	.288	.125	private, claim, jail, removed, banned	-.299	-.224	-.138
cake, christmas, gift, cupcakes, gifts	.330	.287	.125	war, action, army, battle, zone	-.323	-.263	-.135
evening, day, rest, today, sunday	.249	.180	.118	security, transition, administration, support	-.295	-.225	-.134
light, dark, colors, bright, rainbow	.244	.178	.114	general, major, impact, signs, conflict	-.295	-.227	-.132
shopping, home, spend, packed, grocery	.326	.301	.111	largest, launches, announces, lands, add	-.273	-.196	-.132
dreams, live, forget, remember, along	.247	.194	.107	guns, planes, riot, weapons, soldiers	-.251	-.165	-.131
darling, xo, hugs	.259	.211	.106	title, tech, stats, division, technical	-.314	-.258	-.129
brother, mom, daddy, daughter, sister	.302	.275	.105	breaking, turns, breaks, falls, puts	-.266	-.190	-.128
moment, awkward, laugh, excitement, laughter	.282	.247	.103	million, billion	-.277	-.206	-.128
totally, awesome, favorite, love, fave	.272	.233	.103	steve, joe, dave, larry, phil	-.294	-.236	-.124
breakfast, dinner, lunch, cooking, meal	.280	.245	.103	football, pitch, blues, derby, lineup	-.276	-.211	-.124
makeup, glasses, lipstick	.264	.223	.102	ceo, warren	-.240	-.160	-.123
Unigrams	**Perc**	**Real**	**Cont**	**Unigrams**	**Perc**	**Real**	**Cont**
love,my,so,!,you,I,her,hair,feel,today,	.339	.259	.156	game,the,sports,against,football,teams,	-.270	-.236	-.130
friends,baby,cute,girls,beautiful,me,heart,			↓	player,fans,report,team,ebola,vs,nba,games,			↓
little,shopping,happy,because,wonderful,				economy,score,government,ceo,americans,			
gorgeous,bed,clothes,am,have,yay,your	.179	.081	.071	goals,app,penalties,play,shit,political,war	-.117	-.062	-.065
Emotion	**Perc**	**Real**	**Cont**	**Emotion**	**Perc**	**Real**	**Cont**
Joy	.255	.245	.091	Anger	-.156	-.117	-.076
				Fear	-.183	-.145	-.084

Table 3: Textual features highlighting errors in human perception of gender compared to ground truth labels. Table shows correlation to perceived gender expression (**Perc**), to ground truth (**Real**) and to perceived gender expression controlled for ground truth (**Cont**). All correlations of gender unigrams, topics and emotions are statistically significant at p < .001 (t-test)

Gender – High Confidence				**Gender – Low Confidence**			
Topic	Conf	Real	Cont	Topic	Conf	Real	Cont
sibling,flirted,married,husband,wife	(.028)	(.071)	.240	wiser,easier,shittier,happier,worse	-.277	(.081)	-.295
fellaz,boyss,dayz,girlz,gurlz,sistas	(.118)	(.113)	.221	agenda,planning,activities,schedule	-.285	(.020)	-.289
brother, mom, daddy, daughter, sister	(.127)	.241	.214	horoscope,zodiac,gemini,taurus,virgo	-.269	(.087)	-.288
bathroom,wardrobe,toilet,clothes,bath	(.017)	.220	.212	reshape,enable,innovate,enhance,create	-.253	(-.110)	-.235
looked,winked,smiled,lol'd,yell,stare	(.035)	(.089)	.201	imperfect,emotional,break-down,commit	-.227	.024	-.232
hair, blonde, shave, eyebrows, dye	.163	.182	.199	major,brief,outlined,indicates,wrt	-.234	(-.045)	-.226
pyjama,shirt,coat,hoody,trousers	(.077)	(-.010)	.191	justification,circumstance,boundaries	-.224	(-.014)	-.221
awake, dream, sleep, asleep, nights	.160	(.132)	.184	experiencing, explanations, expressive	-.225	(-.039)	-.217
totally, awesome, favorite, love, fave	(.063)	(.135)	. 183	inferiority,sufficiently,adequately	-.209	(-.015)	-.206
days,minutes,seconds,years,months	(.087)	(-.013)	.177	specified,negotiable,exploratory,expert	-.190	(-.014)	-.187
baldy,gangster,boy,kid,skater,dude	(.071)	(.027)	.173	multiple,desirable,extensive,increasingly	-.199	(-.092)	-.183
shopping,grocery,ikea,manicure	(.052)	.204	.173	anticipate,optimist,unrealistic,exceed	(.053)	(.023)	-.182
happy,birthdayyyy,happyyyy,bday	.180	.222	.172	organisation,communication,corporate	-.200	-.148	-.175
girl, lucky, she's, you're, he's	(.118)	(.060)	.172	hostile,choppy,chaotic,cautious,neutral	-.178	(-.033)	-.172
worst,happiest,maddest,slowest,funniest	.173	(.113)	.172	security, transition, administration, supports	.185	(-.079)	-.170
bazillion,shitload,nonstop,spent,aand	.162	(.084)	.167	diminished,unemployment,rapidly	-.181	(-.101)	-.163
Emotion	**Conf**	**Real**	**Cont**	**Emotion**	**Conf**	**Real**	**Cont**
Joy	.202	.245	.164	–			
Anticipation	.140	(.086)	.124				
Unigrams	**Conf**	**Real**	**Cont**	**Unigrams**	**Conf**	**Real**	**Cont**
I,my,this,was,me,so,had,like,	.312	.267	.360	more,may,might,although,	.290	.081	.310
her,night,she,just,hair,gonna,			↓	emotional,your,eager,url,			↓
ever,last,shirt,				desires,relationship,seem,existing,			
kid,girls,love	(.076)	(.047)	.160	emotions,surface,practical,source	.150	-.014	.180

Table 4: Textual features highlighting high and low confidence in human perception of gender. Table shows correlation to average self-reported confidence (**Conf**), to ground truth (**Real**) and with self-reported confidence controlled for ground truth (**Cont**). All correlations of gender unigrams, topics and emotions are statistically significant at p < .001 (t-test), except of the values in brackets.

Perceived – Older				**Perceived – Younger**			
Topic	**Perc**	**Real**	**Cont**	**Topic**	**Perc**	**Real**	**Cont**
golf, sport, semi, racing	.278	(.085)	.226	she's, youre, hes, lucky, girl, slut	-.328	-.243	-.184
bill, union, gov, labor, cuts	.349	.287	.181	boys, girls, hella, homies, ya'll	-.297	-.236	-.155
states, public, towns, area, employees, immigrants	.301	.213	.173	dumb, petty, weak, lame, bc, corny	-.295	-.232	-.155
roger, stanley, captain	.232	(.105)	.167	miss, doing, chilling, how's	-.305	-.268	-.145
available, service, apply, package, customer	.279	.197	.160	heart, cry, smile, deep, hug	-.258	-.186	-.144
serving, prime, serve, served, freeze	.215	(.097)	.154	friend, bestfriend, boyfriend, bff, bestest	-.281	-.254	-.127
support, leaders, group, youth, educate	.228	.121	.153	ugly, stubborn, bein, rude, childish, greedy	-.238	-.182	-.126
hillary, clinton, obama, president, scott, ed, sarah	.289	.230	.150	bitch, fuck, hoe, dick, slap, suck	-.278	-.251	-.125
via, daily, press, latest, report, globe	.311	.272	.149	kinda, annoying, weird, silly, emo, retarded, random	-.242	-.193	-.124
diverse, developed, multiple, among, several, highly	.266	.195	.147	everyone, everything, nothing, does, anyone, else	-.201	-.218	-.118
military, terrorist, citizens, iraq, refugees	.287	.235	.146	bruh, aye, fam, doin, yoo, dawg	-.227	-.178	-.117
julia, emma, annie, claire	.180	(.056)	.145	ever, cutest, worst, weirdest, biggest, happiest	-.275	-.264	-.115
liberty, pacific, north, eastern, 2020	.260	.198	.139	seriously, crazy, bad, shitty, yikes, insane	-.208	-.152	-.114
brooklyn, nyc, downtown, philly, hometown	.213	.120	.139	whoops, oops, remembered, forgot	-.179	(-.104)	-.113
Unigrams	**Perc**	**Real**	**Cont**	**Unigrams**	**Perc**	**Real**	**Cont**
golf, our, end, delay, favourite, low, holes, original,	.321	.(063)	.282	me, i, when, like, you, so, dude, don't, hate, im, u,	-.535	-.489	-.294
branch, the, of, stanley, our, . , story, , ,				girl, hate, life, my, wanna, literally,			
forever, exciting, great, what, community, hurricane,			↓	r, really, cute, someone, youre, miss, me , want, this			↓
for, brands, toward, kids, regarding, upcoming	.208	(.101)	.145	okay, rt, school, snapchat, shit, crying	-.256	(-.051)	-.117
Emotion	**Perc**	**Real**	**Cont**	**Emotion**	**Perc**	**Real**	**Cont**
Positive	.325	.268	.166	Disgust	-.177	-.131	-.094
Trust	.243	.184	.130	Negative	-.104	(-.031)	-.084
Anticipation	.212	.176	.102	Sadness	-.126	-.072	-.081
				Anger	-.070	(-.009)	-.065

Table 5: Textual features highlighting errors in human perception of age compared to ground truth labels. Table shows correlation to perceived age expression (**Perc**), to ground truth (**Real**) and to perceived age expression controlled for ground truth (**Cont**). All correlations of age unigrams, topics and emotions are statistically significant at $p < .001$ (t-test), except of the values in brackets.

Age – High Confidence				**Age – Low Confidence**			
Topic	**Conf**	**Real**	**Cont**	**Topic**	**Conf**	**Real**	**Cont**
school, student, college, teachers, grad, classroom	.242	(-.054)	.227	mocho, gracias, chicos, corazon, quiero	-.195	(-.042)	-.207
done, homework, finished, essay, procrastinating	.251	-.125	.219	sweepstakes, giveaway, enter, retweet, prize	(-.044)	-.278	-.134
math, chem, biology, test, study, physics	.227	(-.060)	.210	injures, shot, penalty, strikes, cyclist, suffered	-.149	.153	-.108
cant, can't, wait, till, believe, afford	.226	-.171	.183	final, cup, europa, arsenal, match, league	-.135	.107	-.106
tomorrow, friday, saturday, date, starts	.175	(-.014)	.171	juventus, munich, lyon, bayern, 0-1	(-.101)	(-.005)	-.103
invitations, prom, attire, wedding, outfit, gowns	.172	(.005)	.170	castlevania, angels, eagles, demons, flames	-.138	.138	(-.101)
soexcited, next, week, weekend, summer, graduation	.153	(.009)	.155	devil, sword, curse, armor, die, obey	(-.081)	(-.055)	(-.097)
aaand, after, before, literally, off, left, gettingold	.182	(-.103)	.154	football, reds, kickoff, derby, pitch, lineup	-.125	.106	(-.096)
sleepy, work, shifts, longday, exhausted, nap	.126	(.064)	.144	anime, invader, shock, madoka, dragonball	(-.071)	(-.080)	(-.095)
life, daydream, remember, cherish, eternally, reminiscing	.200	-.228	.143	paranormal, dragon, alien, zombie, dead	(-.099)	(.025)	(-.092)
happyyyy, birthdaaaay, b-day, bday, belated	.187	-.173	.142	earthquake, magniture, aftermath, devastating, victims	(-.101)	(.040)	(-.090)
Unigrams	**Conf**	**Real**	**Cont**	**Unigrams**	**Conf**	**Real**	**Cont**
my, i'm, can't, i, school, so, to, class,	.375	-.350	.314	rt, his, league, epic	(-.023)	-.320	-.128
semester, college, homework, prom, me, in my,			↓	warriors, ! ,			↓
friends, literally, when, exam, nap	.180	(.080)	.157	vintage	-.130	(.071)	-.111
Emotion	**Conf**	**Real**	**Cont**	**Emotion**	**Conf**	**Real**	**Cont**
Trust	(.077)	.184	.134	–			
Joy	.125	(.009)	.128				
Positive	(.031)	.268	.115				
Anticipation	(.060)	.176	.114				

Table 6: Textual features highlighting high and low confidence in human perception of age. Table shows correlation to average self-reported confidence (**Conf**), to ground truth (**Real**) and with self-reported confidence controlled for ground truth (**Cont**). Correlation values of age unigrams, topics and emotions statistically significant at $p < .001$ (t-test) unless in brackets.

labels. Table 3 displays the features with significant correlation to perceived gender expression when controlled for real gender using partial correlation, as well as the standalone correlations with the real gender label and perceived gender expression. Note that all correlations with both males and females have the same sign for both perceived gender and real gender. This highlights that humans are *not* wrong in using these features to make gender assessments. Rather, these stereotypical associates are overestimated by humans.

By analyzing the topics that are still correlated with perception after controlling for ground truth correlation, we see that topics related to sports, politics, business and technology are considered by annotators to be stronger cues for predicting males than they really are. Female perception is dominated by topics and words relating to feelings, shopping, dreaming, housework and beauty. For emotions, joy is perceived to be more associated to females than the data shows, while users expressing more anger and fear are significantly more likely to be perceived as males than the data supports.

Our crowdsourcing experiment allowed annotators to self-report their confidence in each choice. This gives us the opportunity to measure which textual features lead to higher self-reported confidence in predicting user traits. Table 4 shows the textual features most correlated with self-reported confidence of the annotators when controlled for ground truth, in order to account for the effect that overall confidence is on average higher for groups of users that are easier to predict (i.e., females in case of gender, younger people in case of age).

Annotations are most confident when family relationships or other people are mentioned, which aid them to easily assign a label to a user (e.g., 'husband'). Other topics leading to high confidence are related to apparel or beauty. Also the presence of joy leads to higher confidence (for predicting females based on the previous result). Low confidence is associated with work related topics or astrology as well as to clusters of general adverbs and verbs and tentatively, to a more formal vocabulary e.g., 'specified', 'negotiable', 'exploratory'. Intriguingly, low confidence in predicting gender is also related to unigrams like 'emotions', 'relationship', 'emotional'.

7.2 Age Perception

Table 5 displays the features most correlated with perceived age – the average of the 9 annotator guesses – when controlled for real age, and the individual correlations to perceived and real age.

Again, annotators relied on correct stereotypes, but relied on them more heavily than warranted by data. The results show that the perception of users as being older compared to their biological age, is driven by topics including politics, business and news events. Vocabulary contains somewhat longer words (e.g., 'regarding', 'upcoming', 'original'). Additionally, annotators perceived older users to express more positive emotions, trust and anticipation. This is in accordance with psychology research, which showed that both positive emotion (Mather and Carstensen, 2005) and trust (Poulin and Haase, 2015) increase as people get older.

The perception of users being younger than their biological age is highly correlated with the use of short and colloquial words, and self-references, such as the personal pronoun 'I'. Remarkably, the negative sentiment is perceived as more specific of younger users, as well as the negative emotions of disgust, sadness and anger, the later of which is actually uncorrelated to age.

Table 6 displays the features with the highest correlation to annotation confidence in predicting age when controlling for the true age, as well as separate correlations to real and perceived age. Annotators appear to be more confident in their guess when the posts display more joy, positive emotion, trust and anticipation words. In terms of topics mentioned, these are more informal, self-referential or related to school or college. Topics leading to lower confidence are either about sports or online contests or are frequently retweets.

8 Conclusions

This is the first study to systematically analyze differences between real user traits and traits as perceived from text, here Twitter posts. Overall, participants were generally accurate in guessing a person's traits supporting earlier research that stereotypical associations are frequently accurate (McCauley, 1995). However, we have demonstrated that humans use stereotypes which lead to systematic biases by comparing their guesses to predictions from statistical models using the bag-of-words assumption. While qualitatively different, these predictions were shown to offer complimentary information in case of gender, boosting overall accuracy when used jointly.

Our experimental design allowed us to directly test which textual cues lead to inaccurate assessments. Correlation analysis showed that aspects of stereotypes associated with errors tended not to be completely wrong but rather poorly applied. Annotators generally exaggerated the diagnostic utility of behaviors that they correctly associated with one group or another. Further, we used the same methodology to analyze self-reported confidence.

Follow-up studies can analyze the perception of other user traits such as education level, race or political orientation. Another avenue of future research can look at the annotators' own traits and how these relate to perception (Flekova et al., 2015). This would allow to uncover demographic or psychological traits that influence the ability to make more accurate judgements. This is particularly useful in offering task requesters a prior over which annotators are expected to perform tasks better.

Acknowledgments

The authors acknowledge the support from Templeton Religion Trust, grant TRT-0048. The work of the first author was further supported by the German Research Foundation under grant No. GU 798/14-1 and the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01-S12054. We wish to thank Prof. Iryna Gurevych for supporting the collaboration, the Mechanical Turk annotators for their diligence and the reviewers for their thoughtful comments.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender Identity and Lexical Variation in Social Media. *Journal of Sociolinguistics*, 18(2):135–160.

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. In *Intelligent Tutoring Systems*, volume 3220, pages 592–603.

Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting Social Power Relationships from Natural Language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 773–782.

Matthias Braunhofer, Mehdi Elahi, and Francesco Ricci. 2015. User personality and the new user problem in a context-aware point of interest recommender system. In *Information and Communication Technologies in Tourism*, pages 537–549.

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1301–1309.

Subhagata Chattopadhyay, Preetisha Kaur, Fethi Rabhi, and Rajendra Acharya. 2011. An Automated System to Diagnose the Severity of Adult Depression. In *Second International Conference on Emerging Applications of Information Technology*, EAIT, pages 121–124.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying Mental Health Signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, ACL, pages 51–60.

Aron Culotta, Nirmal Kumar Ravi, and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data. In *Proceedings of the 9th International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 72–78.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56.

John F Dovidio, John C Brigham, Blair T Johnson, and Samuel L Gaertner. 1996. Stereotyping, Prejudice, and Discrimination: Another Look. *Stereotypes and Stereotyping*, 276:319.

Alice H Eagly. 1995. The Science and Politics of Comparing Women and Men. *American Psychologist*, 50(3):145–158.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1277–1287.

Lucie Flekova and Iryna Gurevych. 2013. Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media - Notebook for PAN at CLEF 2013. In *CLEF 2013 Labs and Workshops - Online Working Notes*.

Lucie Flekova, Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2015. Analyzing Crowdsourced Assessment of User Traits through Twitter Posts. In *Third AAAI Conference on Human Computation and Crowdsourcing*, HCOMP.

Lucie Flekova, Lyle Ungar, and Daniel Preoctiuc-Pietro. 2016. Exploring Stylistic Variation with Age and Income on Twitter. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL.

Goldine C. Gleser, Louis A. Gottschalk, and Watkins John. 1959. The Relationship of Sex and Intelligence to Choice of Words: A Normative Study of Verbal Behavior. *Journal of Clinical Psychology*, 15(2):182–191.

Dirk Hovy and Anders Søgaard. 2015. Tagging Performance Correlates with Author Age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 483–488.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual Syntactic Variation over Age and Gender. In *Proceedings of the 19th Conference on Computational Language Learning*, CONNL, pages 103–112.

Oliver P John and Richard W Robins. 1994. Accuracy and Bias in Self-Perception: Individual Differences in Self-enhancement and the Role of Narcissism. *Journal of Personality and Social Psychology*, 66(1):206–219.

Leda E. Kanellakos. 2002. Formal Vocabulary as a Status Cue: Interactions with Diffuse Status Characteristics.

David A. Kenny and Linda Albright. 1987. Accuracy in Interpersonal Perception: A Social Relations Analysis. *Psychological Bulletin*, 102(3):390–402.

Diane Kobrynowicz and Nyla R Branscombe. 1997. Who Considers themselves Victims of Discrimination?: Individual Difference Predictors of Perceived Gender Discrimination in Women and Men. *Psychology of Women Quarterly*, 21(3):347–363.

Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private Traits and Attributes are Predictable from Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 405–413.

Vasileios Lampos, Nikolaos Aletras, Jens K. Geyti, Bin Zou, and Ingemar J. Cox. 2016. Inferring the Socioeconomic Status of Social Media Users based on Behaviour and Language. In *Proceedings of the 38th European Conference on Information Retrieval*, ECIR, pages 689–695.

A Bryan Loyall and Joseph Bates. 1997. Personality-rich believable agents that use language. In *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 ACL 2012 System Demonstrations*, ACL, pages 25–30.

Mara Mather and Laura L Carstensen. 2005. Aging and Motivated Cognition: The Positivity Effect in Attention and Memory. *Trends in Cognitive Sciences*, 9(10):496–502.

Clark R. McCauley. 1995. Are Stereotypes Exaggerated? A Sampling of Racial, Gender, Academic, Occupational, and Political Stereotypes. *Stereotype accuracy: Toward appreciating group differences*, pages 215–243.

Allen R McConnell and Russell H Fazio. 1996. Women as Men and People: Effects of Gender-marked Language. *Personality and Social Psychology Bulletin*, 22(10).

Julie R. McMillan, A. Kay Clifton, Diane McGrath, and Wanda S. Gale. 1977. Women's language: Uncertainty or interpersonal sensitivity and emotionality? *Sex Roles*, 3(6):545–559.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at the International Conference on Learning Representations*, ICLR, pages 1–12.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Anthony Mulac, Lisa B. Studley, and Sheridan Blau. 1990. The Gender-linked Language Effect in Primary and Secondary Students' Impromptu Essays. *Sex Roles*, 23(9-10):439–470.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3):211–236.

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 7th International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 439–448.

Dong-Phuong Nguyen, RB Trieschnigg, AS Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and FMG 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.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*.

James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. 2003. Psychological Aspects of Natural Language Use: Our Words, our Selves. *Annual Review of Psychology*, 54(1):547–577.

Michael Poulin and Claudia Haase. 2015. Growing to Trust. Evidence That Trust Increases and Sustains Well-Being Across the Life Span. *Social Psychological and Personality Science*, 6(6):614–621.

Daniel Preoţiuc-Pietro, Johannes Eichstaedt, Gregory Park, Maarten Sap, Laura Smith, Victoria Tobolsky, H Andrew Schwartz, and Lyle H Ungar. 2015a. The Role of Personality, Age and Gender in Tweeting about Mental Illnesses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, NAACL, pages 21–30.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015b. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015c. 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.

Kimberly Quinn and Neil Macrae. 2005. Categorizing Others: The Dynamics of Person Construal. *Journal of Personality and Social Psychology*, 88(3):467–479.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd Author Profiling Task at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, CLEF.

Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd Author Profiling Task at PAN 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum (Working Notes)*, CLEF.

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*, SMUC, pages 37–44.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1650–1659.

Rinat B Rosenberg-Kima, Amy L Baylor, E Ashby Plant, and Celeste E Doerr. 2008. Interface Agents as Social Models for Female Students: The Effects of Agent Visual Presence and Appearance on Female Students' Attitudes and Beliefs. *Computers in Human Behavior*, 24(6):2741–2756.

Sara Rosenthal and Kathleen McKeown. 2011. Age Prediction in Blogs: A Study of Style, Content, and Online Behavior in pre- and post-Social Media Generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 763–772.

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.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of Age and Gender on Blogging. AAAI Spring Symposium, pages 199–205.

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(9).

Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti. 2015. Your Installed Apps Reveal your Gender and More! *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(3):55–61.

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86:420–428.

Robert Tibshirani. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288.

Svitlana Volkova and Yoram Bachrach. 2015. On Predicting Sociodemographic Traits and Emotions from Communications in Social Networks and Their Implications to Online Self-Disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18(12):726–736.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring Perceived Demographics from User Emotional Tone and User-Environment Emotional Contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring User Political Preferences from Streaming Communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 186–196.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring Latent User Properties from Texts Published in Social Media. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (Demo)*, AAAI, pages 4296–4297.

Wu Youyou, Michal Kosinski, and David Stillwell. 2015. Computer-based Personality Judgments are more Accurate than those Made by Humans. *PNAS*, 112(4):1036–1040.