Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment

Dong Nguyen[14*] **Dolf Trieschnigg**[14] **A. Seza Doğruöz**[23] **Rilana Gravel**[4]
Mariët Theune[1] **Theo Meder**[4] **Franciska de Jong**[1]
(1) Human Media Interaction, University of Twente, Enschede, The Netherlands
(2) Netherlands Institute for Advanced Studies, Wassenaar, NL
(3) Tilburg School of Humanities, Tilburg University, Tilburg, NL
(4) Meertens Institute, Amsterdam, The Netherlands
*Corresponding author: d.nguyen@utwente.nl

Abstract

There is a growing interest in automatically predicting the gender and age of authors from texts. However, most research so far ignores that language use is related to the social identity of speakers, which may be different from their biological identity. In this paper, we combine insights from sociolinguistics with data collected through an online game, to underline the importance of approaching age and gender as social variables rather than static biological variables. In our game, thousands of players guessed the gender and age of Twitter users based on tweets alone. We show that more than 10% of the Twitter users do not employ language that the crowd associates with their biological sex. It is also shown that older Twitter users are often perceived to be younger. Our findings highlight the limitations of current approaches to gender and age prediction from texts.

1 Introduction

A major thrust of research in sociolinguistics aims to uncover the relationship between social variables such as age and gender, and language use (Holmes and Meyerhoff, 2003; Eckert and McConnell-Ginet, 2013; Eckert, 1997; Wagner, 2012). In line with scholars from a variety of disciplines, including the social sciences and philosophy, sociolinguists consider age and gender as social and fluid variables (Eckert, 2012). Gender and age are shaped depending on the societal context, the culture of the speakers involved in a conversation, the individual experiences and the multitude of social roles: a female teenager might also be a high school student, a piano player, a swimmer, etc. (Eckert, 2008).

Speakers use language as a resource to construct their identity (Bucholtz and Hall, 2005). For example, a person's gender identity is constructed through language by using linguistic features associated with male or female speech. These features gain social meaning in a cultural and societal context. On Twitter, users construct their identity through interacting with other users (Marwick and boyd, 2011). Depending on the context, they may emphasize specific aspects of their identity, which leads to linguistic variation both within and between speakers. We illustrate this with the following three tweets:

Tweet 1: *I'm walking on sunshine <3 #and don't you feel good*
Tweet 2: *lalaloveya <3*
Tweet 3: *@USER loveyou ;D*

In these tweets, we find linguistic markers usually associated with females (e.g. a heart represented as <3). Indeed, 77% of the 181 players guessed that a female wrote these tweets in our online game. However, this is a 16-year old biological male, whose Twitter account reveals that he mostly engages with female friends. Therefore, he may have accommodated his style to them (Danescu-Niculescu-Mizil et al., 2011) and as a result he employs linguistic markers associated with the opposite biological sex.

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1950–1961, Dublin, Ireland, August 23-29 2014.

Most of the NLP research focusing on predicting gender and age has approached these variables as *biological* and *static*, rather than *social* and *fluid*. For example, current approaches use supervised machine learning models trained on tweets from males and females. However, the resulting stereotypical models are ineffective for Twitter users who tweet differently from what is to be expected from their biological sex.

As explained above, language use is based on social gender and age identity, and not on biological sex and chronological age. In other words, treating gender and age as fixed biological variables in analyzing language use is too simplistic. By comparing the *biological sex and chronological age* of Twitter users with how they are perceived by the crowd (as an indication of socially constructed identities), we shed light on the *difficulty* of predicting gender and age from language use and draw attention to the *inherent limitations* of current approaches.

As has been demonstrated in several studies, the crowd can be used for experimentation (e.g., Munro et al. (2010)). Our study illustrates the value of the crowd for the study of human behavior, in particular for the experimental study of the social dimension of language use. To collect data, we created an online game (an example of *gamification* (Deterding et al., 2011)) in which thousands of players (the crowd) guessed the biological sex and age of Twitter users based on only the users' tweets. While variance between annotators has traditionally been treated as noise, more recently variation is being treated as a *signal* rather than noise (Aroyo and Welty, 2013). For example, Makatchev and Simmons (2011) analyze how English utterances are perceived differently across language communities.

This paper follows this trend, treating variation as meaningful information. We assume that the crowd's perception (based on the distribution of the players' guesses) is an indication of to what extent Twitter users emphasize their gender and age identity in their tweets. For example, when a large proportion of the players guess the same gender for a particular user, the user is assumed to employ linguistic markers that the crowd associates with gender-specific speech (e.g. iconic hearts used by females).

Our contributions are as follows:

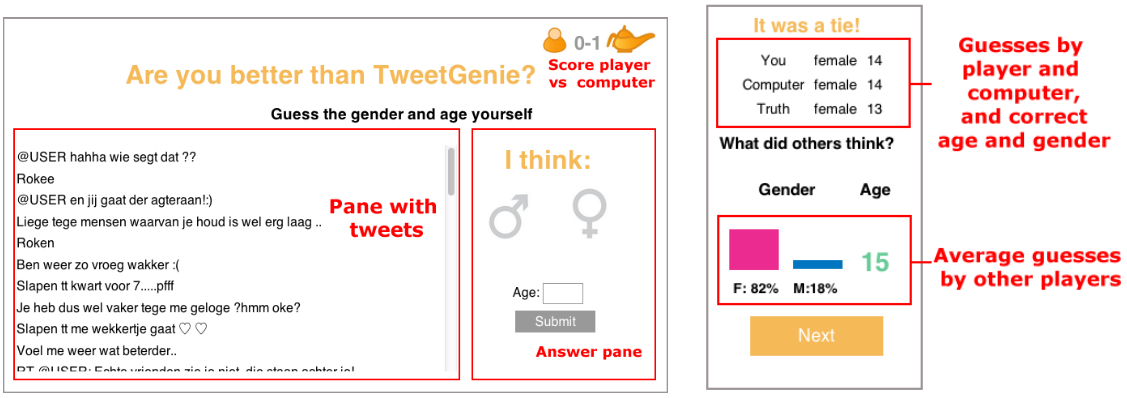
- We demonstrate the use of gamification to study sociolinguistic research problems (Section 3).
- We study the difficulty of predicting an author's gender (Section 4) and age (Section 5) from text alone by analyzing prediction performance by the crowd. We relate our results to sociolinguistic theories and show that approaching gender and age as fixed biological variables is too simplistic.
- Based on our findings, we reflect on current approaches to predicting age and gender from text, and draw attention to the limitations of these approaches (Section 6).

2 Related Work

Gender Within sociolinguistics, studies on gender and language have a long history (Eckert and McConnell-Ginet, 2013). More recently, the NLP community has become increasingly interested in this topic. Most of the work aims at predicting the gender of authors based on their text, thereby focusing more on prediction performance than sociolinguistic insights.

A variety of datasets have been used, including Twitter (Rao et al., 2010; Bamman et al., 2014; Fink et al., 2012; Bergsma and Van Durme, 2013; Burger et al., 2011), blogs (Mukherjee and Liu, 2010; Schler et al., 2005), telephone conversations (Garera and Yarowsky, 2009), YouTube (Filippova, 2012) and chats in social networks (Peersman et al., 2011). Females tend to use more pronouns, emoticons, emotion words, and blog words (*lol, omg, etc.*), while males tend to use more numbers, technology words, and links (Rao et al., 2010; Bamman et al., 2014; Nguyen et al., 2013). These differences have also been exploited to improve sentiment classification (Volkova et al., 2013) and cyberbullying detection (Dadvar et al., 2012).

To the best of our knowledge, the study by Bamman et al. (2014) is the only computational study that approaches gender as a social variable. By clustering Twitter users based on their tweets, they show that multiple gendered styles exist. Unlike their study, we use the crowd and focus on implications for gender and age prediction.


0-1
Score player vs computer
Are you better than TweetGenie?
Guess the gender and age yourself
Pane with tweets
I think:
♂
♀
Age:
Submit
Answer pane
It was a tie!
Guesses by player and computer, and correct age and gender
What did others think?
Gender
Age
15
F: 82%
M:18%
Average guesses by other players
Next

@USER hahha wie segt dat ??
Rokee
@USER en jij gaat der agteraan!:)
Liege tege mensen waarvan je houd is wel erg laag ..
Roken
Ben weer zo vroeg wakker :(
Slapen tt kwart voor 7.....pfff
Je heb dus wel vaker tege me geloge ?hmm oke?
Slapen tt me wekkertje gaat ♡ ♡
Voel me weer wat beterder..
RT @USER: Echte vrienden zie je niet, die staan achter je!

You	female	14
Computer	female	14
Truth	female	13

Figure 1: Screenshot of the game. Text is translated into English (originally in Dutch). Left shows the interface when the user needs to make a guess. Right shows the feedback interface.

Age Eckert (1997) makes a distinction between chronological (number of years since birth), biological (physical maturity), and social age (based on life events). Most of the studies on language and age focus on chronological age. However, speakers with the same chronological age can have very different positions in society, resulting in variation in language use. Computational studies on language use and age usually focus on automatic (chronological) age prediction. This has typically been modeled as a classification problem, although this approach often suffers from ad hoc and dataset dependent age boundaries (Rosenthal and McKeown, 2011). In contrast, recent works also explored predicting age as a continuous variable and predicting lifestages (Nguyen et al., 2013; Nguyen et al., 2011) .

Similar to studies on gender prediction, a variety of resources have been used for age prediction, including Twitter (Rao et al., 2010; Nguyen et al., 2013), blogs (Rosenthal and McKeown, 2011; Goswami et al., 2009), chats in social networks (Peersman et al., 2011) and telephone conversations (Garera and Yarowsky, 2009). Younger people use more alphabetical lengthening, more capitalization of words, shorter words and sentences, more self-references, more slang words, and more Internet acronyms (Rosenthal and McKeown, 2011; Nguyen et al., 2013; Rao et al., 2010; Goswami et al., 2009; Pennebaker and Stone, 2003; Barbieri, 2008).

3 Data

To study how people perceive the gender and age identity of Twitter users based on their tweets, we created an online game. Players were asked to guess the gender and age of Twitter users from tweets. The game was part of a website (TweetGenie, www.tweetgenie.nl) that also hosted an automatic system that predicts the gender and age of Twitter users based on their tweets (Nguyen et al., 2014). To attract players, a link to the game was displayed on the page with the results of the automatic prediction, and visitors were challenged to test if they were better than the automatic system (TweetGenie).

3.1 Twitter Data

We sampled Dutch Twitter users in the fall of 2012. We employed external annotators to annotate the biological sex and chronological age (in years) using all information available through tweets, the Twitter profile and external social media profiles such as Facebook and Linkedin. In total over 3000 Twitter users were annotated. For more details regarding the collection of the dataset we refer to Nguyen et al. (2013).

We divided the data into train and test sets. 200 Twitter users were randomly selected from the test set to be included in the online game (statistics are shown in Table 1). Named entities were manually anonymized to conceal the user's identity. Names in tweets were replaced by 'similar' names (e.g. a first name common in a certain region in the Netherlands was replaced with another common name in that region). This was done without knowing the actual gender and age of the Twitter users. Links were replaced with a general [LINK] token and user mentions with @USER.

Gender and age	F, <20	M, <20	F, [20-40)	M, [20-40)	F, ≥40	M, ≥40
Frequency	61	60	24	23	17	15

Table 1: Statistics Twitter users in our game

3.2 Online Game

Game Setup The interface of the game is shown in Figure 1. Players guessed the biological sex (male or female) and age (years) of a Twitter user based on only the tweets. For each user, {20, 25, 30, 35, 40} tweets were randomly selected. For a particular Twitter user, the same tweets were displayed to all players. Twitter users were randomly selected to be displayed to the players.

To include an entertainment element, players received feedback after each guess. They were shown the correct age and gender, the age and gender guessed by the computer, and the average guessed age and gender distribution by the other players. In addition, a score was shown of the player versus the computer.

Collection In May 2013, the game was launched. Media attention resulted in a large number of visitors (Nguyen et al., 2014). We use the data collected from May 13, 2013 to August 21, 2013, resulting in a total of 46,903 manual guesses. Players tweeted positively about the game, such as '*@USER Do you know what is really addictive? "Are you better than Tweetgenie" ...*' and '*@USER Their game is quite fun!*' (tweets translated to English).

We filter sessions that do not seem to contain genuine guesses: when the entered age is 80 years or above, or 8 or below. These thresholds were based on manual inspection, and chosen because it is unlikely that the shown tweets are from users of such ages. For each guess, we registered a session ID and an IP address. A new session started after 2 hours of inactivity. To study player performance more robustly, we excluded multiple sessions of the same player. After three or more guesses had been made in a session, all next sessions from the same IP address were discarded.

Statistics Statistics of the data are shown in Table 2. Figure 2 shows the distribution of the number of guesses per session. The longest sessions consisted of 18 guesses. Some of our analyses require multiple guesses per player. In that case, we only include players having made at least 7 guesses.

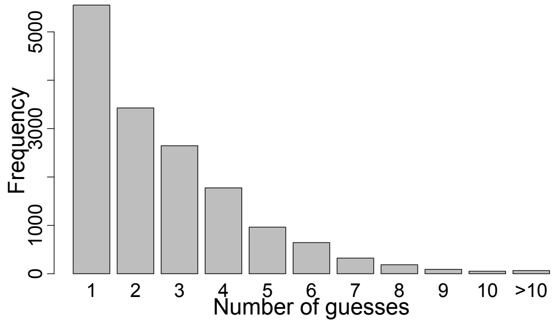


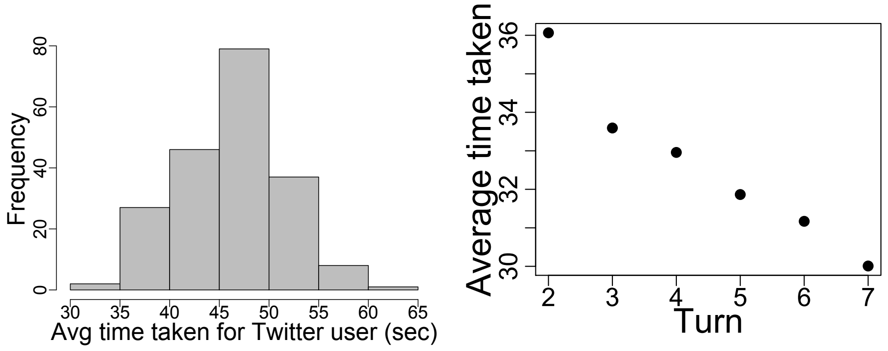
Figure 2: Number of guesses per session

# guesses	41,989
# sessions	15,724
Avg. time (sec) per guess	46
Avg. # guesses / session	2.67

Table 2: Statistics online game (after cleaning)

We calculate the time taken for a guess by taking the time difference between two guesses (therefore, no time for the first guess in each session could be measured). For each Twitter user, we calculate the average time that was taken to guess the gender and age of the user. (Figure 3a). There is a significant correlation (Pearson's r = 0.291, $p < 0.001$) between the average time the players took to evaluate the tweets of a Twitter user and the number of displayed tweets.

There is also a significant correlation between the average time taken for a user and the entropy over gender guesses (Pearson's r = 0.410, $p < 0.001$), and the average time taken for a user and the standard deviation of the age guesses (Pearson's r = 0.408, $p < 0.001$). Thus, on average, players spent more time on Twitter users for whom it was more difficult to estimate gender and age.



(a) Average time taken for Twitter users

(b) Average time taken per turn

Figure 3: Time taken in game

We observe that as the game progresses, players tend to take less time to make a guess. This is shown in Figure 3b, which shows the average time taken for a turn (restricted to players with at least 7 guesses). There was no significant correlation between time spent on a guess and the performance of players and we did not find trends of performance increase or decrease as players progressed in the game.

3.3 Automatic Prediction

Besides studying human performance, we also compare the predictions of humans with those of an automatic system. We split the data into train and test sets using the same splits as used by Nguyen et al. (2013). We train a logistic regression model to predict gender (male or female), and a linear regression model to predict the age (in years) of a person.

More specifically, given an input vector $\mathbf{x} \in \mathbb{R}^m$, $x_1, \ldots, x_m$ represent features. In the case of gender classification (e.g. $y \in \{-1, 1\}$), the model estimates a conditional distribution $P(y|\mathbf{x}, \beta) = 1/(1 + exp(-y(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})))$, where β_0 and $\boldsymbol{\beta}$ are the parameters to estimate. Age is treated as a regression problem, and we find a prediction $\hat{y} \in \mathbb{R}$ for the exact age of a person $y \in \mathbb{R}$ using a linear regression model: $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. We use Ridge (also called L_2) regularization to prevent overfitting.

We make use of the liblinear (Fan et al., 2008) and scikit-learn (Pedregosa et al., 2011) libraries. We only use unigram features, since they have proven to be very effective for gender (Bamman et al., 2014; Peersman et al., 2011) and age (Nguyen et al., 2013) prediction. Parameters were tuned using cross-validation on the training set.

4 Gender

Most of the computational work on language and gender focuses on gender classification, treating gender as fixed and classifying speakers into females and males. However, this assumes that gender *is* fixed and is something people have, instead of something people *do* (Butler, 1990).

In this section, we first analyze the *task difficulty* by studying crowd performance on inferring gender from tweets. We observe a relatively large group of Twitter users who employ language that the crowd associates with the *opposite* biological sex. This, then, raises questions about the upper bound that a prediction system based on only text can achieve.

Next, we place Twitter users on a *gender continuum* based on the guesses of the players and show that treating gender as a binary variable is too simplistic. While historically gender has been treated as binary, researchers in fields such as sociology (Lorber, 1996) and sociolinguistics (Holmes and Meyerhoff, 2003; Bergvall et al., 1996) find this view too limited. Instead, we assume the simplest extension beyond a binary variable: a one-dimensional gender continuum (or scale) (Bergvall et al., 1996). For example, Bergvall (1999) talks about a '*continuum of humans' gendered practices*'. While these previous studies were based on qualitative analyses, we take a quantitative approach using the crowd.

4.1 Task Difficulty

Majority vote We study crowd performance using a system based on the *majority* of the players' guesses. Majority voting has proven to be a strong baseline to aggregate votes (e.g. in crowdsourcing systems (Snow et al., 2008; Le et al., 2010)). On average, we have 210 guesses per Twitter user, providing substantial evidence per Twitter user. A system based on majority votes achieves an accuracy of 84% (Table 3a shows a confusion matrix). Table 3b shows a confusion matrix of the majority predictions versus the automatic system. We find that the biological sex was predicted incorrectly by both the majority vote system and the automatic system for 21 out of the 200 Twitter users (10.5%, not in Table).

Automatic classification systems on English tweets achieve similar performances as our majority vote system (e.g. Bergsma and Van Durme (2013) report an accuracy of 87%, Bamman et al. (2014) 88%). More significantly, the results suggest that 10.5% (automatic + majority) to 16% (majority) of the Dutch Twitter users do not employ language that the crowd associates with their biological sex. As said, this raises the question of whether we can expect much higher performances by computational systems based on only language use.

		Biological sex	
		Male	Female
Crowd	Male	82	16
	Female	16	86

(a) Crowd (majority)

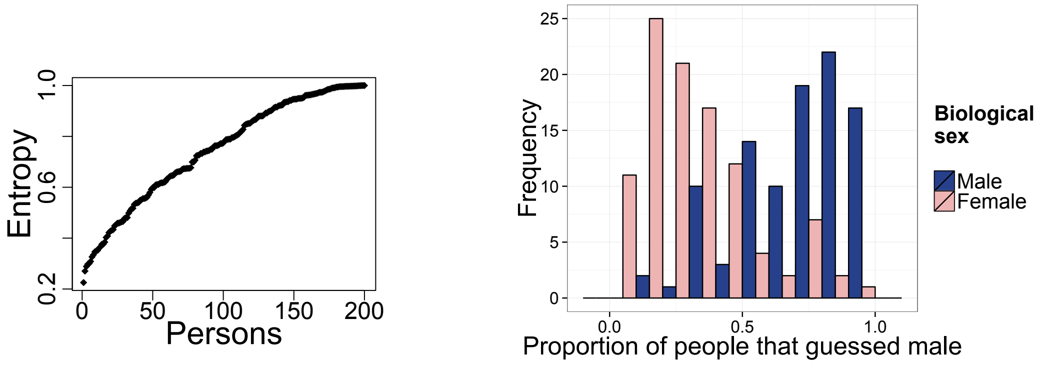
		Crowd	
		Male	Female
Automatic	Male	68	22
	Female	30	80

(b) Automatic vs crowd

Table 3: Confusion matrices crowd prediction

Individual players versus an automatic system When considering players with 7 or more guesses, the average accuracy for a player is 0.71. Our automatic system achieves an accuracy of 0.69. The small number of tweets per Twitter user in our data (20-40) makes it more difficult to automatically predict gender.

Entropy We characterize the difficulty of inferring a user's gender by calculating the entropy for each Twitter user based on the gender guesses (Figure 4a). We find that the difficulty varies widely across users, and that there are no distinct groups of 'easy' and 'difficult' users. However, we do observe an interaction effect between the entropy of the gender guesses and the ages of the Twitter users. At an aggregate level, we find no significant trend. Analyzing females and males separately, we observe a significant trend with females (Pearson's r = 0.270, $p < 0.01$), suggesting that older female Twitter users tend to emphasize other aspects than their gender in tweets (as perceived by the crowd).



(a) Entropy over gender guesses

(b) A histogram of all Twitter users and the proportion of players who guessed the users were male. For example, there are 25 female users for which 10 - 20% of the players guessed they were male.

Figure 4: Gender prediction

4.2 Binarizing Gender, a Good Approach?

Using data collected through the online game we *quantitatively* put speakers on a gender continuum based on how their tweets are perceived by the crowd. For each Twitter user, we calculate the proportion of players who guessed the users were male and female. A plot is displayed in Figure 4b. We can make the following observations:

First, the guesses by the players are based on their expectations about what kind of behaviour and language is used by males and females. The plot shows that for some users, almost all players guessed the same gender, indicating that these expectations are quite strong and that there are stylistic markers and topics that the crowd strongly associates with males or females.

Second, if treating gender as a binary variable is reasonable, we would expect to see two distinct groups. However, we observe quite an overlap between the biological males and females. There are 1) users who conform to what is expected based on their biological sex, 2) users who deviate from what is expected, 3) users whose tweets do not emphasize a gender identity or whose tweets have large variation using language associated with both genders. We investigated whether this is related to their use of Twitter (professional, personal, or both), but the number of Twitter users in our dataset who used Twitter professionally was small and not sufficient to draw conclusions.

We now illustrate our findings using examples. The first example is a 15-year old biological female for who the crowd guessed most strongly that she is female (96% of n=220). Three tweets from her are shown below. She uses language typically associated with females, talking about spending time with her girlfriends and the use of stylistic markers such as hearts and alphabetical lengthening. Thus, she conforms strongly to what the crowd expects from her biological sex.

Tweet 4: *Gezellig bij Emily en Charlotte.*
Translation: *Having fun with Emily and Charlotte.*

Tweet 5: *Hiiiiii schatjesss!*
Translation: *Hiiiiii cutiesss!*

Tweet 6: ♥ *@USER*

Below are two tweets from a 40 year old biological female who does not employ linguistic markers strongly associated with males or females. Therefore, only 46% of the crowd (n=200) was able to guess that she is female.

Tweet 7: *Ik viel op mijn bek. En het kabinet ook. Geinig toch? #Catshuis*
Translation: *I went flat on my face. And the cabinet as well. Funny right? #Catshuis*

Tweet 8: *Jeemig. Ik kan het bijna niet volgen allemaal.*
Translation: *Jeez. I almost can't follow it all.*

Twitter users vary in how much they emphasize their gender in their tweets. As a result, the difficulty of inferring gender from tweets varies across persons, and treating gender as a binary variable ignores much of the interesting variation within and between persons.

Automatic system We now analyze whether an automatic system is capable of capturing the position of Twitter users on the gender continuum (as perceived by the crowd). We calculate the correlation between the proportion of male guesses (i.e. the position on the gender continuum) and the scores of the logistic regression classifier: $\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. While the training data was binary (users were labeled as male or female), a reasonable Spearman correlation of $\rho = 0.584$ (p $<$ 0.001) was obtained between the classifier score and the score based on the crowd's perception. We did not observe a significant relation between the score of the classifier (corresponding to the confidence of the gender prediction) and age.

5 Age

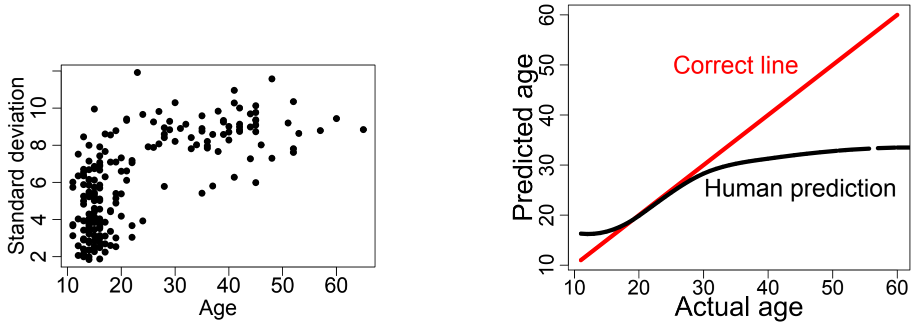
We start with an analysis of task difficulty, by studying crowd performance on inferring age from tweets. Next, we show that it is particularly hard to accurately infer the chronological age of older Twitter users from tweets.

5.1 Task Difficulty

The crowd's average guesses As with a system based on majority vote for gender prediction, we test the performance of a system that predicts the ages of Twitter users based on the average of all guesses. We find that such a system achieves a Mean Absolute Error (MAE) of 4.844 years and a Pearson's correlation of 0.866. Although the correlation is high, the absolute errors are quite large. We find that the crowd has difficulty predicting the ages of older Twitter users. There is a positive correlation (Pearson's ρ = 0.789) between the absolute errors and the actual age of Twitter users. There is a negative correlation between the errors (predicted - actual age) and the actual age of Twitter users (Pearson's ρ = -0.872).

We calculate the standard deviation over all the age guesses for a user (Figure 5a) to measure the difficulty of inferring a user's age. There is a positive correlation between age and standard deviation of the guesses ($\rho = 0.691$), which indicates that players have more difficulty in guessing the ages of older Twitter users.

Individual players versus an Automatic System To estimate the performance of individual players, we restrict our attention to players with at least 7 guesses. We find that individual players are, on average, 5.754 years off. A linear regression system achieves a MAE of 6.149 years and a Pearson correlation of 0.812. The small number of tweets in our data (20-40) increases the difficulty of the task for automatic systems.



(a) Standard deviation and actual age

(b) Average age prediction by humans.

Figure 5: Age prediction

5.2 Inferring the Age of Older Twitter Users

Figure 5b shows the average player predictions with the actual age of the Twitter users. The red line is the 'perfect' line, i.e. the line when the predictions would match the exact age. Black represents a fitted LOESS curve (Cleveland et al., 1992) based on the human predictions. We find that the players tend to overpredict the age of younger Twitter users, but even more strikingly, on average they consistently underpredict the age of older Twitter users. The prediction errors already start at the end of the 20s, and the gap between actual and predicted age increases with age.

This could be explained by sociolinguistic studies that have found that people between 30 and 55 years use standard forms the most, because they experience the maximum societal pressure in the workplace to conform (Holmes, 2013). On Twitter, this has been observed as well: Nguyen et al. (2013) found fewer linguistic differences between older age groups than between younger age groups. This makes it difficult for the crowd to accurately estimate the ages of older Twitter users. Younger people and retired people use more non-standard forms (Holmes, 2013). Unfortunately, our dataset does not contain enough retired users to analyze whether this trend is also present on Twitter.

6 Discussion

We now discuss the implications of our findings for research on automatically predicting the gender and age of authors from their texts.

Age and gender as *social* variables Most computational research has treated gender and age as fixed, biological variables. The dominant approach is to use supervised machine learning methods to generalize across a large number of examples (e.g. texts written by females and males). While the learned models so far are effective at predicting age and gender of *most* people, they learn stereotypical behaviour and therefore provide a simplistic view.

First, by using the crowd we have shown that Twitter users emphasize their gender and age in varying degrees and in different ways, so that for example, treating gender as a binary variable is too simplistic (Butler, 1990; Eckert and McConnell-Ginet, 2013). Many users do not employ the stereotypical language associated with their biological sex, making models that take a static view of gender ineffective for such users. More detailed error analyses of the prediction systems will increase understanding of the reasons for incorrect predictions, and shed light on the relation between language use and social variables.

Second, models that assume static variables will not be able to model the interesting variation (Eisenstein, 2013). Models that build on recent developments in sociolinguistics will be more meaningful and will also have the potential to contribute to new sociolinguistic insights. For example, modeling what influences speakers to show more or less of their identity through language, or jointly modeling variation between and within speakers, are in our opinion interesting research directions. The ever increasing amounts of social media data offer opportunities to explore these research directions.

Sampling We have shown that the difficulty of tasks such as gender and age prediction varies across persons. Therefore, creating datasets for such tasks requires maximum attention. For example, when a dataset is biased towards people who show a strong gender identity (e.g. by sampling followers of accounts highly associated with males or females, such as sororities (Rao et al., 2010)), the results obtained on such a set may not be representative of a more random set (as observed when classifying political affiliation (Cohen and Ruths, 2013)).

Task difficulty Our study also raises the question of what level of performance can be obtained for tasks such as predicting gender and age from only language use. Since we often form an impression based on someone's writing, crowd performance is a good indicator of the task difficulty. While the crowd performance does not need to be the upper bound, it does indicate that it is difficult to predict gender and age of a large number of Twitter users.

When taking the majority label, only 84% of the users were correctly classified according to their biological sex. This suggests that about 16% of the Dutch Twitter users do not use language that the *crowd* associates with their biological sex.

We also found that it is hard to accurately estimate the ages of older Twitter users, and we related this to sociolinguistics studies who found less linguistic differences in older age groups due to societal pressure in the workplace.

Limitations A limitation of our work is that we focused on language variation *between* persons, and not on variation *within* persons. However, speakers vary their language depending on the context and their conversation partners (e.g. accommodation effects were found in social media (Danescu-Niculescu-Mizil et al., 2011)). For example, we assigned Twitter users an overall 'score' by placing them on a gender continuum, ignoring the variation we find within users.

Crowdsourcing as a tool to understand NLP tasks Most research on crowdsourcing within the NLP community has focused on how the crowd can be used to obtain fast and large amounts of annotations. This study is an example of how the crowd can be used to obtain a deeper understanding of an NLP task. We expect that other tasks where disagreement between annotators is meaningful (i.e. it is not only due to noise), could potentially benefit from crowdsourcing experiments as well.

7 Conclusion

In this paper, we demonstrated the successful use of the crowd to study the relation between language use and social variables. In particular, we took a closer look at inferring gender and age from language using data collected through an online game. We showed that treating gender and age as fixed variables ignores the variety of ways people construct their identity through language.

Approaching age and gender as *social* variables will allow for richer analyses and more robust systems. It has implications ranging from how datasets are created to how results are interpreted. We expect that our findings also apply to other social variables, such as ethnicity and status. Instead of only focusing on performance improvement, we encourage NLP researchers to also focus on what we can *learn* about the relation between language use and social variables using computational methods.

Acknowledgements

This research was supported by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO), grants IB/MP/2955 (TINPOT) and 640.005.002 (FACT). The third author is supported through the Digital Humanities research grant by Tilburg University and a NIAS research fellowship. The authors would like to thank the players of the TweetGenie game.

References

Lora Aroyo and Chris Welty. 2013. Crowd truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. In *Proceedings of WebSci'13*.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Federica Barbieri. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics*, 12(1):58–88.

Shane Bergsma and Benjamin Van Durme. 2013. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 710–720.

Victoria L. Bergvall, Janet M. Bing, and Alice F. Freed. 1996. *Rethinking Language and Gender Research: Theory and Practice*. Routledge.

Victoria L. Bergvall. 1999. Toward a comprehensive theory of language and gender. *Language in society*, 28(02):273–293.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

John D. Burger, John 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.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

William S. Cleveland, Eric Grosse, and William M. Shyu. 1992. Local regression models. *Statistical models in S*, pages 309–376.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy! In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pages 91–99.

Maral Dadvar, Franciska de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*, pages 23–25.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World Wide Web*, pages 745–754.

Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Penelope Eckert. 1997. Age as a sociolinguistic variable. *The handbook of sociolinguistics*, pages 151–167.

Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics*, 12(4):453–476.

Penelope Eckert. 2012. Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41:87–100.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 359–369.

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

Katja Filippova. 2012. User demographics and language in an implicit social network. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1478–1488.

Clayton Fink, Jonathon Kopecky, and Maksym Morawski. 2012. Inferring gender from the content of tweets: A region specific example. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 710–718.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Proceedings of the Third International ICWSM Conference*, pages 214–217.

Janet Holmes and Miriam Meyerhoff. 2003. *The handbook of language and gender*. Wiley-Blackwell.

Janet Holmes. 2013. *An introduction to sociolinguistics*. Routledge.

John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. 2010. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 21–26.

Judith Lorber. 1996. Beyond the binaries: Depolarizing the categories of sex, sexuality, and gender*. *Sociological Inquiry*, 66(2):143–160.

Maxim Makatchev and Reid Simmons. 2011. Perception of personality and naturalness through dialogues by native speakers of American English and Arabic. In *Proceedings of the SIGDIAL 2011: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 286–293.

Alice E. Marwick and danah boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 122–130.

Dong Nguyen, Noah A Smith, and Carolyn P. Rosé. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 115–123.

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*, pages 439–448.

Dong Nguyen, Dolf Trieschnigg, and Theo Meder. 2014. Tweetgenie: Development, evaluation, and lessons learned. In *Proceedings of COLING 2014*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 37–44.

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.

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.

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*, pages 763–772.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2005. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pages 199–205.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

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*, pages 1815–1827.

Suzanne E. Wagner. 2012. Age grading in sociolinguistic theory. *Language and Linguistics Compass*, 6(6):371–382.