"How Old Do You Think I Am?": A Study of Language and Age in Twitter

Dong Nguyen[1], Rilana Gravel[2], Dolf Trieschnigg[1], Theo Meder[2]
[1]University of Twente, Enschede, The Netherlands
[2]Meertens Institute, Amsterdam, The Netherlands
{d.nguyen,d.trieschnigg}@utwente.nl,{gravel.rilana,theo.meder}@meertens.knaw.nl

Abstract

In this paper we focus on the connection between age and language use, exploring age prediction of Twitter users based on their tweets. We discuss the construction of a fine-grained annotation effort to assign ages and life stages to Twitter users. Using this dataset, we explore age prediction in three different ways: classifying users into age categories, by life stages, and predicting their exact age. We find that an automatic system achieves better performance than humans on these tasks and that both humans and the automatic systems have difficulties predicting the age of older people. Moreover, we present a detailed analysis of variables that change with age. We find strong patterns of change, and that most changes occur at young ages.

Introduction

A person's language use reveals much about their social identity. A person's social identity is based on the groups he or she belongs to, including groups based on age, gender and political affiliation. Earlier research in sociolinguistics regarded male and female, and age as biological variables. Examples for this are Labov (1966) and Trudgill (1974). However, current research views them primarily as social variables. Concepts such as gender and age are shaped differently depending on an individual's experiences and personality, and the society and culture a person is part of (Eckert 1997; Holmes and Meyerhoff 2003). To complicate things even more, the two variables gender and age are intertwined: studying one of the variables implies studying the other one, as well. For example, the appropriate age for cultural events often differs for males and females (Eckert 1997). Besides linguistic variation based on the groups a person belongs to, there is also variation within a single speaker as people adapt their language to their audience (Bell 1984). Thus it follows that speakers can choose to show gender and age identity more or less explicitly in language use, depending on people's perception of these variables, on their culture, the recipient of their utterance, etc. From a sociolinguistic perspective, language is a resource which can be drawn on to study different aspects of a person's social identity at different points in an interaction (Holmes and Meyerhoff 2003).

Early sociolinguistic studies only had access to relatively small datasets (e.g. a couple of hundred persons), due to time and practical constraints on the collection of data. With the rise of social media such as Twitter, new resources have emerged that can complement these analyses. Compared to previously used resources in sociolinguistic studies like face-to-face conversations, Twitter is interesting in that it collapses multiple audiences into a single context: tweets can be targeted to a person, a group, or to the general public (Marwick and Boyd 2011). Twitter offers the opportunity to gather large amounts of informal language from many individuals. However, the Twitter population might be biased and only little is known about the studied persons. To overcome this, we carried out a large annotation effort to annotate the gender and age of Twitter users. While gender is one of the most studied variables, the relation between age and language has only recently become a topic of interest.

In this paper we present work on automatically predicting people's age. This can offer new insights into the relation between language use and age. Such a system could also be used to improve targeting of advertisements and to support fine-grained analyses of trends on the web. So far, age prediction has primarily been approached by classifying persons into age categories. We revisit this approach being the first to approach age prediction from three different angles: classifying users into *age categories* (20-, 20-40, 40+), predicting their *exact age*, and classifying users by their *life stage* (secondary school student, college student, employee). We compare the performance of an automatic system with that of humans on these tasks. Next, to allow a more fine-grained analysis, we use the exact ages of Twitter users and analyze how language use changes with age.

Specifically, we make the following contributions: 1) We present a characterization of Dutch Twitter users as a result of a fine-grained annotation effort; 2) we explore different ways of approaching age prediction (age categories, life stages and exact age); 3) we find that an automatic system has better performance than humans on the task of inferring age from tweets; 4) we analyze variables that change with age, and find that most changes occur at younger ages.

We start with discussing related work and our dataset. Next, we discuss our experiments on age prediction. We then continue with a more fine-grained analysis of variables that change with age. We conclude with a summary.

Related Work

Eckert (1997) distinguishes between chronological age (number of years since birth), biological age (physical maturity) and social age (based on life events). Studies about language and age usually consider chronological age and apply an *etic* approach, grouping speakers based on age spans (e.g. (Labov 1966; Trudgill 1974; Barbieri 2008)). But speakers can have a very different position in society than their chronological age indicates. Therefore, it might be reasonable to apply an *emic* approach, grouping speakers according to shared experiences of time, such as school as a shared experience for teenagers (Eckert 1997).

So far, automatic age prediction has mostly been approached as a two-class or three-class classification problem based on age spans with for example boundaries at 30 or 40 years (e.g. (Rao et al. 2010; Garera and Yarowsky 2009; Goswami, Sarkar, and Rustagi 2009)), thus corresponding to an etic approach. However, as choosing boundaries still remains problematic, several researchers have looked more closely into this issue. For example, Rosenthal and McKeown (2011) experimented with varying the binary split for creating age categories. In contrast, Nguyen, Smith, and Rosé (2011) approached age prediction as a regression problem, eliminating the need to create age categories. In our work, we will experiment with age prediction as a regression problem, as a classification problem based on age categories and explore an emic approach, by classifying persons according to their life stages.

Both content features and stylistic features (such as part-of-speech and the amount of slang words) have been found to be useful for predicting the age of users (Nguyen, Smith, and Rosé 2011; Argamon et al. 2007; Goswami, Sarkar, and Rustagi 2009). Pennebaker and Stone (2003) found that as people get older, they tend to use more positive and fewer negative words, focus more on the future and less on the past and make fewer self-references. Not much research has been done yet on investigating the relationship between gender and age from a computational perspective. Argamon et al. (2007) found that certain linguistic features that increase with age, also increase more with males. Nguyen, Smith, and Rosé (2011) incorporated gender using a binary variable, only allowing a simple interaction between gender and age. Many others have ignored the effect of gender when predicting the age of users.

Experiments on automatic classification of users according to latent attributes such as gender and age have been done on a wide range of resources, including telephone conversations (Garera and Yarowsky 2009), blogs (Sarawgi, Gajulapalli, and Choi 2011), forum posts (Nguyen, Smith, and Rosé 2011) and scientific articles (Bergsma, Post, and Yarowsky 2012; Sarawgi, Gajulapalli, and Choi 2011). Recently, Twitter has started to attract interest by researchers as a resource to study automatic identification of user attributes, such as ethnicity (Pennacchiotti and Popescu 2011; Rao et al. 2011), gender (Fink, Kopecky, and Morawski 2012; Bamman, Eisenstein, and Schnoebelen 2012; Rao et al. 2010; Burger et al. 2011; Rao et al. 2011), geographical location (Eisenstein et al. 2010) and age (Rao et al. 2010).

Corpus Collection

In this section we describe a large annotation effort we carried out to annotate Dutch Twitter users. Based on the results we present a characterization of Dutch Twitter users.

Selecting and Crawling Users

Twitter users can indicate information such as their name, location, website and short biography in their profile. However, gender and age are not explicit fields in Twitter profiles. As a result, other researchers working on identification of such attributes have resorted to a variety of approaches to construct a corpus, ranging from focused crawling to using lists with common names.

For example, Rao et al. (2010) constructed a corpus by *focused* crawling. To collect users they used a crawl with seeds by looking for profiles that had *'baby boomers', 'junior', 'freshman'* etc. in their description. However, this leads to a potential bias by starting with users that explicitly indicate their age identity in their profile. Burger et al. (2011) sampled users from the Twitter stream and used links to blogging sites, indicated in their profile, to find the gender. Therefore, their set of users was restricted to users having blogs and willing to link them using Twitter. Some approaches used lists of male and female names, for example obtained using Facebook (Fink, Kopecky, and Morawski 2012) or from the US social security department (Zamal, Liu, and Ruths 2012; Bamman, Eisenstein, and Schnoebelen 2012).

Our goal was to select a set of users as randomly as possible, and not biasing user selection by searching on well-known stereotypical behavior or relying on links to explicit sources. This did create the need for a large annotation effort, and resulted in a smaller user sample. Using the Twitter API we collected tweets that contained the word '*het*', which can be used as a definite article or pronoun in Dutch. This allowed us to restrict our tweets to Dutch as much as possible, and limit the risk of biasing the collection somehow. During a one-week period in August 2012 we sampled users according to this method. Of these users, we randomly selected a set for annotation. We then collected all followers and followees of these users and randomly selected additional users from this set. We only included accounts with less than 5000 followers, to limit the inclusion of celebrities and organizations. For all users, we initially downloaded their last 1000 tweets. Then new tweets from these users were collected from September to December 2012.

	Het		**Followe(e/r)s**	
Annotated	1842	(76%)	1343	(43%)
Not enough tweets	15	(0.6%)	129	(4%)
Not a person	221	(9%)	441	(14%)
Not public	264	(11%)	719	(23%)
Not Dutch	51	(2%)	468	(15%)
Other	46	(2%)	17	(0.5%)
Total	2439		3117	

Table 1: Reasons why accounts were discarded/kept by sampling method.

Dutch Twitter Users

In this section we analyze the effect of our sampling procedure, and present a characterization of Dutch Twitter users in our corpus. We employed two students to perform the annotations. Annotations were done by analyzing a user's profile, tweets, and additional external resources (like Facebook or LinkedIn) if available. In this paper, we only focus on the annotations that are relevant to this study.

Effect of Sampling Method

The annotators were instructed to only annotate the users that met the following requirements:

- The account should be publicly accessible.
- The account should represent an actual person (e.g. not an organization).
- The account should have 'sufficient' tweets (at least 10).
- The account should have Dutch tweets (note that this does not eliminate multilingual accounts).

We separated the reasons why accounts were discarded by the two sampling methods (*het* and followers/followees) that were used (the first requirement in the list that was not satisfied was marked). The results are reported in Table 1. We observe that the proportion of actual annotated users is much higher for the users obtained using the query '*het*'. The users obtained by sampling from the followers and followees included more non-Dutch accounts, as well as accounts that did not represent persons. In addition, there was also a group of people who had protected their account between the time of sampling and the time of annotation. In total, 3185 users were annotated.

Gender

The biological gender was annotated for 3166 persons (for some accounts, the annotators could not identify the gender). The gender ratio was almost equal, with 49.5% of the persons being female. However, as we will see later, the ratio depends on age. The annotation of the gender was mostly determined based on the profile photo or a person's name, but sometimes also their tweets or profile description.

Mislove et al. (2011) analyzed the US Twitter population using data from 2006-2009. Using popular female and male names they were able to estimate the gender of 64% of the people, finding a highly biased gender ratio with 72% being male. A more recent study by Beevolve.com however found that 53% were women, based on information such as name and profile.

Age

Because we expected most Twitter users to be young, the following three categories were used: 20-, 20-40, 40+. The age category was annotated for 3110 accounts. The results separated by gender are shown in Table 2[1]. There are more females in the young age group, while there are more men in the older age groups. The same observation was made in statistics reported by Beevolve.com.

[1]Note that this table only takes persons into account for who both age and gender were annotated

	20-	20-40	40+
M	796	488	265
F	1078	316	157

Table 2: Age and gender

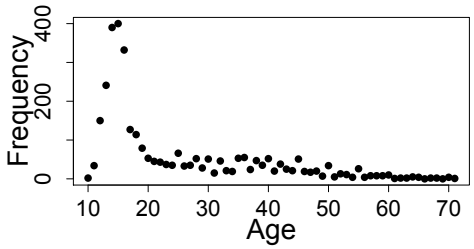

Figure 1: Plot of frequencies per age

We also asked our annotators to annotate the exact age. Sometimes it was possible to get an almost exact estimate, for example by using LinkedIn profiles, exact age mentions in the profile, tweets, or mentioning which grade the person was in. However, since this was not always the case, annotators also indicated a margin (0, 2, 5 or 10 years) of how sure they were. Figure 1 shows a graph with the frequencies per year of age. Table 3 reports the frequencies of the indicated margins. In our data, we find that the margin for young users is low, and that for older users the margin is much higher.

As discussed earlier in this paper, it may be more natural to distinguish users according to their *life stage* instead of a fixed age category. Life stages can be approached from different dimensions. In this paper, we use life stages based on the occupation of people, by distinguishing between students, employed, retired etc. The results are displayed in Table 4. Unfortunately, the decision to annotate this was done while the annotation process was already underway; therefore the accounts of some users were not available anymore (either removed or protected).

We find that the most common life stages are associated with clear age boundaries, although the boundaries are not the same as for the age categories. We find the following age spans in which 90% of the persons fall: secondary school students (12 -16 yrs), college students (16 - 24 yrs), employees (24 - 52 yrs). However, note that with the life stage approach, people may be assigned to a different group than the group that most resembles their age, if this group matches their life stage better. We have plotted the overlap between life stage and age categories in Figure 2.

Age estimation margin	Frequency
0	703
2	1292
5	918
10	173

Table 3: Frequencies of margins for the exact age annotation

Life Stage	Frequency
Secondary school student	1352
College student	316
Employee	1021
Retired	5
Other	15
Unknown	132
Not accessible	344

Table 4: Life stage frequencies

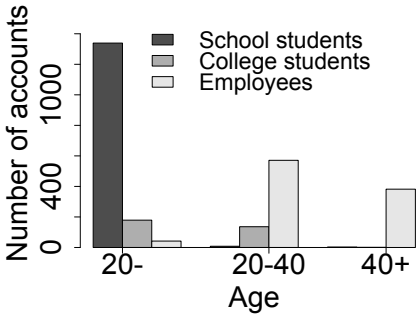

Figure 2: Overlap life stage and age categories

Inter-annotator Agreement

We employed two students to perform the annotations. 84 accounts were annotated by both. Inter-annotator agreement was measured using Cohen's kappa. Generally, a value above 0.7 is considered acceptable. We found the following kappa values: gender (1.0), age category (0.83) and life stage (0.70). For the actual age, the mean absolute difference was 1.59 years.

Age Prediction

Goal

In this section we compare the different ways of approaching age, by testing how feasible age prediction is using simple features based only on the text of tweets. We will automatically predict the following:

- *Age category*: 20-, 20-40, 40+
- *Age*: continuous variable
- *Life stage*: secondary school student[2], college student, employee

For the life stage, we only use categories for which we had a sufficient number of persons. Note that classifying age according to age category and life stage are multiclass classification problems, while treating age as a continuous variable results in a regression problem. In addition, we compare our systems with the performance of humans on this task.

[2]In Dutch this is translated to *scholier*, which includes all students up to and including high school, there is no direct translation in English.

Evaluation

We will evaluate the performance of our classification methods (to predict the age category and life stage) using the F_1 measure. We will report both the macro and micro averages. The regression problem (predicting age as a continuous variable) will be evaluated using the Pearson's correlation coefficient, mean absolute error (MAE) and accuracy, where a prediction was counted as correct if it fell within the margin as specified by the annotators.

Dataset

We restricted our dataset to users who had at least 20 tweets and for whom the gender, age category and exact age were annotated. For each user we sampled up to 200 tweets. We divided the dataset into a train and test set. Each set contains an equal number of males and females, and the same age distribution (according to the annotated age categories) across gender categories. This limits the risks of the model learning features that for example are more associated with a particular gender, due to that gender occuring more in the particular age category. Parameter tuning and development of the features were done using cross-validation on the training set. The statistics are presented in Table 5.

	Train		Test	
	M	**F**	**M**	**F**
20-	602	602	186	186
20-40	231	231	73	73
40+	118	118	37	37
Total	1902		592	

Table 5: Dataset statistics

Learning Algorithm

We use linear models, specifically logistic and linear regression, for our tasks. Given an input vector $\mathbf{x} \in \mathbb{R}^m$, $x_1, \ldots, x_m$ represent features (also called independent variables or predictors). In the case of classification with two classes, e.g. $y\{-1, 1\}$, 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. We use a one versus all method to handle multiclass classification. In the case of regression, 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: $\hat{y} = \beta_0 + \mathbf{x}^\top \boldsymbol{\beta}$. In order to prevent overfitting we use Ridge (also called L_2) regularization. We make use of the liblinear (Fan et al. 2008) and scikit-learn (Pedregosa et al. 2011) libraries.

Preprocessing & Features

Tokenization is done using the tool by (O'Connor, Krieger, and Ahn 2010). All user mentions (e.g. *@user*) are replaced by a common token. Because preliminary experiments showed that a unigram system already performs very well, we only use unigrams to keep the approach simple. We keep words that occur at least 10 times in the training documents. In the next section, we will look at more informed features and how they change as people are older.

Results

In this section we present the results of the three age prediction tasks. The results can be found in Tables 6 and 7. We find that a simple system using only unigram features can already achieve high performance, with micro F_1 scores of above 0.86 for the classification approaches and a MAE of less than 4 years for the regression approach. We also experimented with applying a log transformation of the exact age for the regression task. The predicted values were converted back when calculating the metrics. We find that the MAE and accuracy both improve. In the rest of this section, when referring to the regression run, we refer to the standard run without a log transformation.

Run	F_1 **macro**	F_1 **micro**
Age categories	0.7670	0.8632
Life stages	0.6785	0.8628

Table 6: Results classification

Run	ρ	**MAE**	**Accuracy**
Age regression	0.8845	3.8812	0.4730
Age regression - log	0.8733	3.6172	0.5709

Table 7: Results age regression

A scatterplot of the actual age versus the predicted age can be found in Figure 3. Figure 4 shows the errors per actual age. We find that starting from older ages (around 40-50) the system almost always underpredicts the age. This could have several reasons. It may be that the language changes less as people get older (we show evidence for this in the next section), another plausible reason is that we have very little training data in the older age ranges.

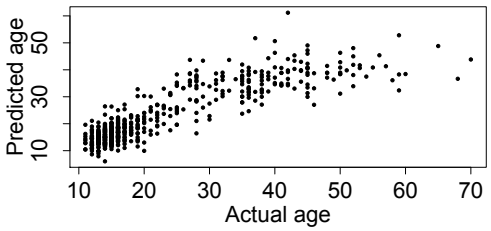

Figure 3: Scatterplot age

The most important features for old and young persons are presented in Tables 8 and 9. We find both content features and stylistic features to be important. For example, content words like *school*, *son*, and *daughter* already reveal much about a person's age. Younger persons talk more about themselves (*I*), and use more chat language such as *haha, xd*, while older people use more conventional words indicating support or wishing well (e.g. *wish, enjoy, thanks, take care*).

For the age categories we redid the classification using only persons for whom the life stage was known to allow better comparison between the two classification tasks. We found that people in the 40+ class are often misclassified as belonging to the 20-40 class, and college students are often classified as secondary school students. The precision and recall for the individual classes are listed in Tables 10 and 11. The performances are comparable. The micro average for life stages is slightly better (0.86 vs 0.85), the macro average is worse (0.68 vs 0.75) as the metric is heavily affected by the bad performance on the *students* class. Although life stages are better motivated from a sociolinguistics viewpoint (Eckert 1997), it is not yet clear which classes are the most suitable. In our corpus, almost all persons were either secondary school students or employees. If a more fine-grained distinction is necessary (for example for personalization), it is still a question which categories should be used.

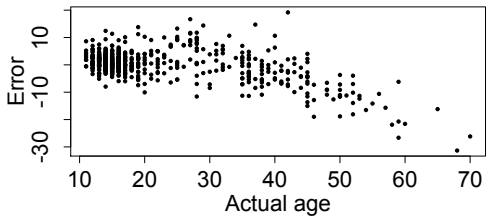

Figure 4: Scatterplot absolute error

Dutch	**English**	**Weight**
school	school	-0.081
ik	I	-0.073
:)	:)	-0.071
werkgroep	work group	-0.069
stages	internships	-0.069
oke	okay	-0.067
xd	xd	-0.066
ben	am	-0.066
haha	haha	-0.064
als	if	-0.064

Table 8: Top features for younger people (regression)

Dutch	**English**	**Weight**
verdomd	damn	0.119
dochter	daugther	0.112
wens	wish	0.112
zoon	son	0.111
mooie	beautiful	0.111
geniet	enjoy	0.110
dank	thanks	0.108
goedemorgen	good morning	0.107
evalueren	evaluate	0.105
sterkte	take care	0.102

Table 9: Top features for older people (regression)

	Precision	**Recall**
20-	0.9297	0.9775
20 - 40	0.6739	0.7561
40+	0.8158	0.4493

Table 10: Results per class: Age categories

	Precision	**Recall**
Sec. school student	0.8758	0.9853
College student	0.6667	0.1250
Employee	0.8541	0.8977

Table 11: Results per class: Life stages

Train	Test	Age categories		Regression			Life stages	
		Macro F1	Micro F1	ρ	MAE	Accuracy	Macro F1	Micro F1
All	F	0.7778	0.8750	0.9101	3.4220	0.5135	0.7038	0.8765
	M	0.7563	0.8514	0.8625	4.3405	0.4324	0.6538	0.8500
Male	F	0.6861	0.8277	0.8784	3.9617	0.5135	0.6151	0.8642
	M	0.7027	0.8311	0.8431	4.5017	0.4459	0.6116	0.8346
Female	F	0.7281	0.8581	0.8965	3.5586	0.5270	0.6438	0.8560
	M	0.6373	0.8041	0.8195	5.2099	0.3682	0.6829	0.8538

Table 12: Effect of gender

Treating age prediction as a regression problem eliminates the need to choose boundaries. The main drawback is that annotating the exact age of users requires more effort than annotating the life stage or an age category. However, as mentioned before, our annotators showed that reliable annotations are possible (on average less than 2 years difference).

In summary, we believe that both classifying users according to their life stage and treating age prediction as a regression problem are promising approaches. Both approaches complement each other. Age prediction as a regression problem relies on chronological age, while life stages are built on shared experiences between people. Depending on the practical application, knowing the chronological age or life stage might be more informative. For example, groups based on life stage might be more useful for marketing purposes, while the chronological age might be more informative when targeting medical information.

Effect of Gender

In Table 12 we have separated the performance according to gender. We also experimented with training on data of only one gender, and reported the performance separated by gender. Across the three tasks the performance for females is better than the performance for males. We also find that across the three tasks, the performance for females is better when trained on only females, compared to the performance of males, when trained on only males.

One of the explanations could be that females write slightly more than men (average #tokens: 2235 versus 2130), although the differences between the means are small and there is no significant difference in the number of tweets per person (note that we sampled up to 200 tweets per person).

Another explanation can be found in sociolinguistic studies. It has been pointed out that females assert their identity more through language than males (Eckert 1989; Labov 1990). Hence, they might use all kinds of in-group vocabulary more often, thereby marking their affiliation with a certain group. Men's vocabulary, on the contrary, is more homogenous across the in-groups (Eckert 2000). Consistent with this, Ling (2005) found that females 'seem to have a broader register when using SMS'. Due to this, it might be easier to determine the age of women. However, neither Eckert (1989) nor Labov (1990) looked at age specifically, and the studied people were also not comparable (e.g. Eckert (1989) only studied young people, and social media settings have not been explored much yet).

Error Analysis

As reported in the previous section, not for all cases the correct age was predicted. This is of course not surprising. People do not only constitute their identity on the basis of their age, but they combine various variables in order to express their selves. For example, a person is not only a teenager, but also a female, a high school student, a piano player, etc. (Eckert 2008). Depending on what a person wants to express at a particular moment and towards a particular person, certain aspects of his/her identity may be more emphasized, making age prediction even more complicated. To illustrate this, we will discuss two Twitter users for whom the age was incorrectly predicted.

Case study 1

The first person is a 24-year old student, who the system estimated to be a 17-year old secondary school student. The top 10 most frequent words for this user are *@USER, RT, •, Ik (I),≪, G, :D, Hahaha, tmi,* and *jij (you)*. The use of special characters like a dot (•) and the much less than sign (≪) is characteristic for younger Twitter users, who separate statements in their tweets employing these characters. *I* is one of the words being the most predictive of younger people as was presented in the feature analysis (see Table 8) and the other words like *hahaha, you* etc. are also highly associated with younger persons in our corpus. As we can see, this person employs these words with such a high frequency that he can easily be mistaken for a secondary school student under 20. Examples containing salient words are the ones below:

@USER kommmdan nurd
@USER comeonthen nurd [nerd]

Hahaahhahaha kkijk rtl gemist holland in da hood, bigga huiltt ik ga stukkk
Hahaahhahaha [I am] wwatching rtl gemist[3] *holland in da hood*[4]*, bigga is cryingg it's killinggg me*

RT @USER: Ook nog eens rennen voor me bus #KutDag • Ik heb weekend :)
RT @USER: Had to run for my bus too #StupidDay • I have weekend :)

In addition to the words mentioned above, *me (my)*, and *heb (have)* appear, which are indicative for younger persons in our corpus, as well.

[3] website where people can watch tv shows online

[4] Dutch reality show

Next to the fact that this person employs words rather associated with teenagers on Twitter, we can also derive what kind of identity is constituted here. In the tweets, unconventional punctuation, emoticons, ellipsis, in-group vocabulary (*nurd*), and alphabetical lengthening (*stukkk*) are used to create an informal, unconventional style particularly addressing an in-group. It can be concluded that this person does not appear to stress his identity as an adult, but finds other aspects of his identity more important to emphasize. These aspects, however, are expressed with features employed most frequently by younger persons in our corpus, resulting in a wrong age prediction for this person.

Case study 2

The second person is a 19-year old student. However, the system predicted him as being a 33-year old employee. The top 10 most frequent words for this user are *@USER, CDA, RT, Ik (I), VVD, SGP, PvdA, D66, bij (at)* and *Groenlinks*. It becomes clear that this person tweets about politics a lot, with Dutch political parties (*CDA, VVD, SGP, D66, Groenlinks*) being six out of his ten most frequent words. Tweets that are characteristic for this user and that relate to some of his most salient words are, for example:

@USER Woensdagochtend 15 augustus start het landelijke CDA met haar regiotour op Goeree-Overflakkee i.s.m. @USER.

@USER On Wednesday morning, the 15th of August the national CDA starts with its tour through the region in Goeree-Overflakkee in collaboration with @USER

RT @USER: Vanmiddag met @USER gezellig bij @USER een wijntje gedaan en naar de Emmaüskerk #Middelharnis geweest. Mooie dag zo!

RT @USER: Had fun this afternoon had wine at @USER with @USER and went to the Emmaüschurch #Middelharnis. Beautiful day!

Almost all of his tweets are (like the first example) about politics, so we can assume this user wants to stress his identity as a person interested in politics, or even as a politician on Twitter. Certainly, this is a more common topic for users older than a 19-year old. Proof for this is the fact that words such as *ministers, elections, voter* etc. are highly ranked features associated with older people in the regression model. In addition, the person uses more prepositions, conventional punctuation, formal abbreviations and for example mentions *wine* which is also rather associated with older people in our corpus. Moreover, *beautiful* is one of the top ten features predictive of older people. Thus, not only the main topic of his tweets (politics) is associated more with older people, but he also represents himself as a grown-up person in his other tweets by using which what we perceive as rather conservative vocabulary and punctuation.

Thus, the discussed cases show that people can emphasize other aspects of identity than age. This can result in a deviation from style and content from their peers, thereby making the automatic prediction of age more difficult.

Manual Prediction

In this section we compare the performance of our systems with the performance of humans on the task of inferring age *only* from tweets. A group of 17 people (including males and females, old and young, active and non-active Twitter users) estimated the gender, life stage, exact age and age categories for a random subset of the Twitter users in the test set. Each person was assigned a different set of about 20 Twitter users. For each Twitter user, a text file was provided containing the same text as used in our automatic prediction experiments. The participants received no additional information such as the name, profile information etc. They could decide themselves how carefully they would read the text, as long as they could make a serious and informed prediction. On average, it took about 60-90 min to do the task. In total there are 337 users for whom we both have manual and automatic predictions. The results can be found in Tables 13 and 14.

Run	F_1 **macro**	F_1 **micro**
Age categories		
Manual	0.619	0.752
Automatic	0.751	0.858
Life stages		
Manual	0.658	0.778
Automatic	0.634	0.853

Table 13: Results classification - manual vs automatic

Run	ρ	**MAE**	**Acc.**
Manual	0.784	4.875	0.552
Automatic	0.879	4.073	0.466

Table 14: Results age regression - manual vs automatic

Using McNemar's Test we find that the automatic system is significantly better in classifying according to age categories (χ^2 = 18.01, df=1, $p < 0.01$) and life stages (χ^2 = 9.76, df=1, $p < 0.01$). The automatic system is also significantly better in predicting the exact age when comparing the MAE's (paired t-test, t(336) = 2.79, $p < 0.01$). In addition, for each metric and task we calculated which fraction of the persons performed *equal or better than* the automatic system. This ranged from 0.24 (age cat., all metrics) to 0.41 (life stages, micro F_1) and 0.47 (life stages, macro F_1), to 0.29 (exact age, MAE's) and 0.82 (exact age, accuracy).

In addition we find the following. First, humans achieve a better accuracy for the regression task. The accuracy is based on margins as indicated by the annotators. Humans were often closer at the younger ages, where the indicated margins were also very low and a slightly off prediction would not be counted as correct. Second, humans have trouble predicting the ages of older people as well. The correlation between the MAE's and exact ages are 0.58 for humans and 0.60 for the automatic system. Third, humans are better in classifying people into life stages than in age categories.

To conclude, we find that an automatic system is capable of achieving *better* performance than humans, and being much faster (on average, taking less than a second compared to 60-90 minutes to predict the age of 20 users).

Variables that change with age

By analyzing the importance of features in an automatic prediction system, only general effects can be seen (i.e. this feature is highly predictive for old versus young). However, to allow for a more detailed analysis, we now use the exact ages of Twitter users to track how variables change with age.

Variables

We explore variables that capture style as well as content.

Style

The following style variables capture stylistic aspects that a person is aware of and explicitly chooses to use:

- *Capitalized words*, for example *HAHA* and LOL. The words need to be at least 2 characters long.
- *Alphabetical lengthening*, for example *niiiiiice* instead of *nice*. Matching against dictionaries was found to be too noisy. Therefore, this is implemented as the proportion of words that have a sequence of the same three characters in the word. The words should also contain more than one unique character (e.g. tokens such as *www* are not included) and contain only letters.
- *Intensifiers*, which enhance the emotional meaning of words (e.g. in English, words like *so, really* and *awful*).

The following variables capture stylistic aspects that a person usually is not aware of:

- *LIWC-prepositions*, the proportion of prepositions such as *for, by* and *on*. The wordlist was obtained from the Dutch LIWC (Zijlstra et al. 2005) and contains 48 words.
- *Word length*, the average word length. Only tokens starting with a letter are taken into account, so hashtags and user mentions are ignored. Urls are also ignored.
- *Tweet length*, the average tweet length.

References

Pennebaker and Stone (2003) found that as people get older, they make fewer self-references. We adapt the categories for the Dutch LIWC (Zijlstra et al. 2005) to use on Twitter data by including alphabetical lengthening, slang, and English pronouns (since Dutch people often tweet in English as well).

- *I*, such as *I, me, mine, ik, m'n, ikke*.
- *You*, such as *you, u, je, jij*.
- *We*, such as *we, our, ons, onszelf, wij*.
- *Other*, such as *him, they, hij, haar*.

Conversation

- *Replies*, proportion of tweets that are a reply or mention a user (and are not a retweet).

Sharing

- *Retweets*, proportion of tweets that are a retweet.
- *Links*, proportion of tweets that contain a link.
- *Hashtags*, proportion of tweets that contain a hashtag.

Variable	Females ρ	Males ρ
Style		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
References		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
Conversation		
Replies	0.304**	0.026
Sharing		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

Table 15: Analysis of variables. For both genders n = 1247. Bonferroni correction was applied to p-values. $^{*}p \leq 0.01$ $^{**}p \leq 0.001$

Analysis

We calculate the Pearson's correlation coefficients between the variables and the actual age using the same data from the age prediction experiments (train and test together), and report the results separated by gender in Table 15.

We find that younger people use more explicit stylistic modifications such as alphabetical lengthening and capitalization of words. Older people tend to use more complex language, with longer tweets, longer words and more prepositions. Older people also have a higher usage of links and hashtags, which can be associated with information sharing and impression management. The usage of pronouns is one of the variables most studied in relation with age. Consistent with Pennebaker and Stone (2003) and Barbieri (2008) we find that younger people use more first-person (e.g. *I*) and second person singular (e.g. *you*) pronouns. These are often seen as indicating interpersonal involvement. In line with the findings of (Barbieri 2008), we also find that older people more often use first-person plurals (e.g. *we*).

In Figure 5 we have plotted a selection of the variables as they change with age, separated by gender. We also show the fitted LOESS curves (Cleveland, Grosse, and Shyu 1992). One should keep in mind that we have less data in the extremes of the age ranges. We find strong changes in the younger ages; however after an age of around 30 most variables show little change. What little sociolinguistics research there is on this issue has looked mostly at individual features. Their results suggest that the differences between age groups above age 35 tend to become smaller (Barbieri 2008). Such trends have been observed with stance (Barbieri 2008) and tag questions (Tottie and Hoffmann 2006). Related to this, it has been shown that adults tend to be more conservative in their language, which could also explain the observed trends. This has been attributed to the pressure of using standard language in the workplace in order to be taken seriously and get or retain a job (Eckert 1997).

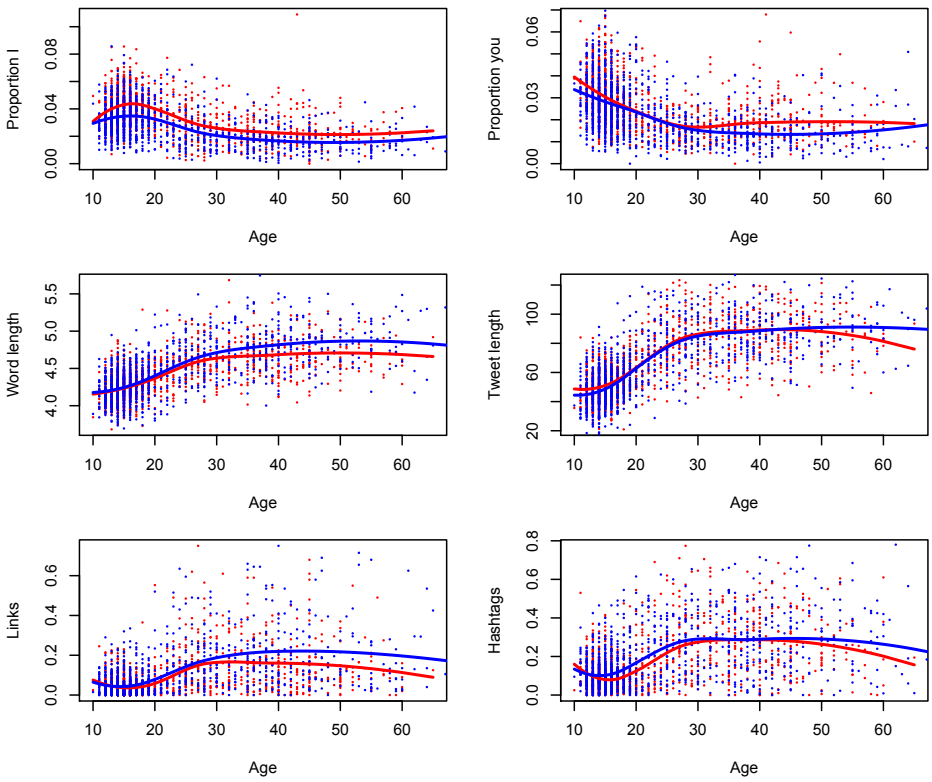

Figure 5: Plots of variables as they change with age. Blue: males, Red: females

One should keep in mind, however, that we have studied people with different ages, and we did not perform a longitudinal study that looked at changes within persons as they became older. Therefore the observed patterns may not indicate actual change within persons, but could be a reflection of changes between different generations.

Reflecting on the age prediction task and the analysis presented in this section, we make the following observations. First, for some variables there is almost no difference between males and females (e.g. tweet length), while for some other variables one of the genders consistently uses that variable more (e.g. the first singular pronouns for females, links for men). In our prediction experiments, we also observed differences in the prediction performance between genders. We also found differences in the gender distribution across age categories on Twitter. Therefore, we conclude that researchers interested in the relation between language use and age should not ignore the gender variable.

Second, in the automatic prediction of exact age we found that as people get older the system almost always underpredicts the age. When studying how language changes over time, we find that most change occurs in the younger ages, while at the older ages most variables barely change. This could be an explanation of why it is harder to predict the correct age of older people (for both humans and the automatic system). This also suggests that researchers wanting to improve an automatic age prediction system should focus on improving prediction for older persons, and thus identifying variables that show more change at older ages.

Conclusion

We presented a study on the relation between the age of Twitter users and their language use. A dataset was constructed by means of a fine-grained annotation effort of more than 3000 Dutch Twitter users. We studied age prediction based only on tweets. Next, we presented a detailed analysis of variables as they change with age.

We approached age prediction in different ways: predicting the age category, life stage, and the actual age. Our system was capable of predicting the exact age within a margin of 4 years. Compared with humans, the automatic system performed better and was much faster than humans. For future research, we believe that life stages or exact ages are more meaningful than dividing users based on age groups. In addition, gender should not be ignored as we showed that how age is displayed in language is also strongly influenced by the gender of the person.

We also found that most changes occur when people are young, and that after around 30 years the studied variables show little change. This may also explain why it is more difficult to predict the age of older people (for both humans and the automatic system).

Our models were based *only* on the tweets of the user. This has as a practical advantage that the data is easy to collect, and thus the models can easily be applied to new Twitter users. However, a deeper investigation into the relation between language use and age should also take factors such as the social network and the direct conversation partners of the tweeters into account.

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 authors would like to thank Mariët Theune and Leonie Cornips for feedback, Charlotte van Tongeren and Daphne van Kessel for the annotations, and all participants of the user study for their time and effort.

References

Argamon, S.; Koppel, M.; Pennebaker, J.; and Schler, J. 2007. Mining the blogosphere: age, gender, and the varieties of self-expression. *First Monday* 12(9).

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2012. Gender in Twitter: styles, stances, and social networks. *CoRR*.

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

Beevolve.com. An exhaustive study of Twitter users across the world. http://www.beevolve.com/twitter-statistics/. Last accessed: Jan 2013.

Bell, A. 1984. Language style as audience design. *Language in society* 13(2):145–204.

Bergsma, S.; Post, M.; and Yarowsky, D. 2012. Stylometric analysis of scientific articles. In *NAACL 2012*.

Burger, J. D.; Henderson, J.; Kim, G.; and Zarrella, G. 2011. Discriminating gender on Twitter. In *EMNLP 2011*.

Cleveland, W.; Grosse, E.; and Shyu, W. 1992. Local regression models. *Statistical models in S* 309–376.

Eckert, P. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

Eckert, P. 1997. *Age as a sociolinguistic variable*. The handbook of sociolinguistics. Blackwell Publishers.

Eckert, P. 2000. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Wiley-Blackwell.

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

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2010. A latent variable model for geographic lexical variation. In *EMNLP 2010*.

Fan, R.-E.; Chang, K.-W.; Hsieh, C.-J.; Wang, X.-R.; and Lin, C.-J. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9.

Fink, C.; Kopecky, J.; and Morawski, M. 2012. Inferring gender from the content of tweets: A region specific example. In *ICWSM 2012*.

Garera, N., and Yarowsky, D. 2009. Modeling latent biographic attributes in conversational genres. In *ACL-IJCNLP 2009*.

Goswami, S.; Sarkar, S.; and Rustagi, M. 2009. Stylometric analysis of bloggers' age and gender. In *ICWSM 2009*.

Holmes, J., and Meyerhoff, M. 2003. *The handbook of language and gender.* Oxford: Blackwell.

Labov, W. 1966. *The social stratification of English in New York City*. Centre for Applied Linguistics.

Labov, W. 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change* 2(2):205–254.

Ling, R. 2005. The sociolinguistics of SMS: An analysis of SMS use by a random sample of Norwegians. *Mobile Communications* 335–349.

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

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. In *ICWSM 2011*.

Nguyen, D.; Smith, N. A.; and Rosé, C. P. 2011. Author age prediction from text using linear regression. In *LaTeCH 2011*.

O'Connor, B.; Krieger, M.; and Ahn, D. 2010. TweetMotif: exploratory search and topic summarization for Twitter. In *ICWSM 2010*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to Twitter user classification. In *ICWSM 2011*.

Pennebaker, J., and Stone, L. 2003. Words of wisdom: Language use over the life span. *Journal of personality and social psychology* 85(2):291.

Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in Twitter. In *SMUC 2010*.

Rao, D.; Paul, M.; Fink, C.; Yarowsky, D.; Oates, T.; and Coppersmith, G. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *ICWSM 2011*.

Rosenthal, S., and McKeown, K. 2011. Age prediction in blogs: a study of style, content, and online behavior in pre- and post-social media generations. In *ACL 2011*.

Sarawgi, R.; Gajulapalli, K.; and Choi, Y. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *CoNLL 2011*.

Tottie, G., and Hoffmann, S. 2006. Tag questions in British and American English. *Journal of English Linguistics* 34(4):283–311.

Trudgill, P. 1974. *The social differentiation of English in Norwich*. Cambridge University Press.

Zamal, F. A.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: inferring latent attributes of Twitter users from neighbors. In *ICWSM 2012*.

Zijlstra, H.; van Middendorp, H.; van Meerveld, T.; and Geenen, R. 2005. Validiteit van de Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC). *Netherlands Journal of Psychology* 60(3).