TweetGenie: Automatic Age Prediction From Tweets

Dong Nguyen
University of Twente
and
Rilana Gravel
Meertens Institute
and
Dolf Trieschnigg
University of Twente
and
Theo Meder
Meertens Institute

A person's language use reveals much about the person's social identity, which is based on the social categories a person belongs to including age and gender. We discuss the development of TweetGenie, a computer program that predicts the age of Twitter users based on their language use. We explore age prediction in three different ways: classifying users into age categories, by life stages, and predicting their exact age. An automatic system achieves better performance than humans on these tasks. Both humans and the automatic systems tend to underpredict the age of older people. We find that most linguistic changes occur when people are young, and that after around 30 years the studied variables show little change, making it difficult to predict the ages of older Twitter users.

DOI: 10.1145/2528272.2528276 http://doi.acm.org/10.1145/2528272.2528276

1. INTRODUCTION

Reading someone's tweets, one often gets a glimpse of the identity of the user. For example, what kind of person would you expect behind the following tweet? [1]

> AS LONG AS YOU LOVE ME <3

And what about this one?

> Interesting article about usability design on mobile search [LINK]

[1] The first example is from a 13 year old female, the second example is from a 30 year old male.

As you can see, a person's language use reveals much about the person's social identity, which is based on the social categories a person belongs to including age and gender. 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]. 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, due to time and practical constraints on the collection of data. Social media, like Twitter, offer the opportunity to gather large amounts of informal language from many individuals. By approaching the problem as a prediction task, researchers have become interested in developing computational systems that predict the gender and age of Twitter users based on their language use. Often, new sociolinguistic insights are obtained when developing such systems.

In May 2013 we launched an online demo called TweetGenie (http://www.tweetgenie.nl). TweetGenie can predict the age and gender of Dutch Twitter users (with a public account) based on the 200 most recent tweets. This paper summarizes the development of TweetGenie, with a focus on age prediction. We refer to [Nguyen et al. 2013] for more details.

Studies about language and age usually consider chronological age, grouping speakers based on age spans [Labov 1966; Trudgill 1974; Barbieri 2008]. Similarly, age prediction has primarily been approached by classifying persons into age categories for example with boundaries at 30 or 40 years (e.g. [Rao et al. 2010; Garera and Yarowsky 2009; Goswami et al. 2009]). But speakers can have a very different position in society than their chronological age indicates. Therefore, it might be reasonable to to group speakers into life stages according to 'shared experiences of time', such as school for teenagers [Eckert 1997].

We revisit the age prediction task being the first to approach age prediction from three different angles: classifying users into *age categories* (20-, 20-40, 40+), classifying users by their *life stage* (secondary school student, college student, employee), and predicting their *exact age*. In this paper we discuss the dataset, the age prediction experiments, and why it is hard to predict the age of older Twitter users.

2. DATASET

The development of TweetGenie started with constructing a dataset. We limited our focus to Dutch Twitter user and sampled Dutch Twitter users based on a commonly used Dutch word (*het*, which can be used as a definite article or pronoun). This restricted the sample to Dutch users and limited a bias towards certain user groups or topics. Twitter users can add information such as their name, location, website and short biography in their profile. However, gender and age are not explicit fields in Twitter profiles. We employed two students to annotate the sample of Dutch Twitter users. Their annotations were based on the users' tweets, the Twitter profile or information from other publicly available social network profiles such as LinkedIn and Facebook. Various properties were annotated, such as the gender, the age category, the life stage and the exact age. More than 3000 users were annotated and annotations and tweets were collected in fall 2012.

3. AGE PREDICTION

3.1 Task setup

We will automatically predict the:

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

	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 I. Dataset statistics

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. The statistics are presented in Table I.

3.2 Learning the Model

We use linear models, specifically logistic and linear regression, for our tasks. In order to prevent overfitting we use Ridge (also called L_2) regularization. Tokenization is done using the tool by [O'Connor et al. 2010]. All user mentions (e.g. *@user*) are replaced by a common token. As features we only use unigrams (single words), because these were found to already perform well in preliminary experiments.

3.3 Results

We find that a system using only unigram features achieves high performance, with micro F_1 scores of above 0.86 for the classification approaches and an average Mean Absolute Error of less than 4 years for the regression approach. A scatterplot of the actual age versus the predicted age can be found in Figure 1. 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 limited training data in the older age ranges.

The most important features for younger and older persons are presented in Table II. Both content features and stylistic features are important. Younger persons talk more about themselves (*I*), and use more chat language such as *haha* and *xd*, while older people use more conventional words indicating support or wishing well (e.g. *wish, enjoy* and *thanks*).

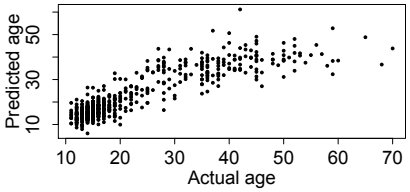

Fig. 1. Scatterplot age predictions

Young	school (school), ik (I), :), werkgroep (work group), stages (internships), oke (okay), xd, ben (am), haha
Old	verdomd (damn), dochter (daughter), wens (wish), zoon (son), mooie (beautiful), geniet (enjoy), dank (thanks)

Table II. Top features (regression)

4. ERROR ANALYSIS

Because a person's identity is not only based on age, but also on other social variables, it is not surprising that the system was not correct in all cases. 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. We illustrate this using an example.

The person is a 24-year old student, who the system estimated to be a 17-year old secondary school student. This person frequently uses special characters like a dot (•) and the much less than sign (≪), both characteristic for younger Twitter users, who separate statements in their tweets employing these characters. In addition, most frequently used words of this person include words like *I, hahahaha, you* etc. that are highly associated with younger persons in our corpus. Example tweets are given below:

> Hahaahhahaha kkijk rtl gemist holland in da hood, bigga huiltt ik ga stukkk
> *Hahaahhahaha [I am] wwatching rtl gemist*[2] *holland in da hood*[3]*, 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 the tweets, unconventional punctuation, emoticons, ellipsis, and alphabetical lengthening (*stukkk*) are used to create an informal, unconventional style particularly addressing an in-group. 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 an incorrect age prediction for this person.

We compared the performance of the system with that of humans. 17 persons participated and each of them had 20 Twitter users assigned. They were asked to guess the age of the Twitter users when only seeing the tweets. The results revealed that an automatic system achieves slightly better (but significant) performance than humans on inferring age from tweets. Humans also needed a significant amount of time (60-90 min.) to do the task. This confirms that predicting the age from language alone is not trivial. Similarly to computers, humans also underpredict the ages of older Twitter users.

5. WHY IS IT HARD TO PREDICT THE AGE OF OLDER PEOPLE?

Based on our error analysis and results we found that the system often underpredicts the age of older Twitter users. Feedback from visitors of TweetGenie confirmed this as well.

For a variety of variables, including the use of pronouns, tweet length and word length we

[2] website where people can watch tv shows online
[3] Dutch reality show

find strong changes in the younger ages; however after an age of around 30 most variables show little change.

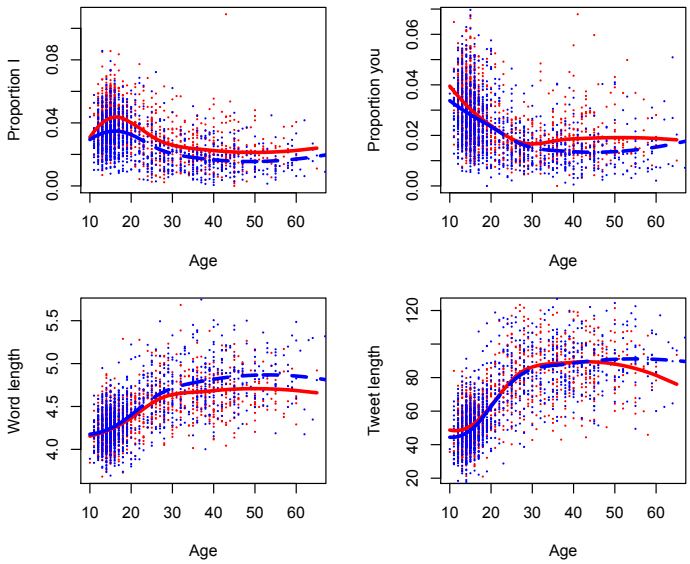

Fig. 2. Plots of variables as they change with age. Blue: males, Red: females

In Figure 2 we have plotted a selection of the variables as they change with age, separated by gender. We also show the fitted LOESS curves [Cleveland et al. 1992]. 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]. This is a possible explanation of why it is harder to predict the correct age of older people.

6. CONCLUSION

In this paper we summarized the development of TweetGenie, a computer program that predicts the age of Twitter users based on their language use [Nguyen et al. 2013]. We approached age prediction from three different angles: classifying users into age categories, predicting their exact age, and classifying users by their life stage. We found that most linguistic 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). As discussed in the error analysis, the way a person speaks is influenced by many other factors besides a person's age. As future work, we plan to do more fine-grained analyses that also take factors such as the social network and the direct conversation partners of the tweeters into account.

7. 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

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

CLEVELAND, W., GROSSE, E., AND SHYU, W. 1992. Local regression models. *Statistical models in S*, 309–376.

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

ECKERT, P. 2008. Variation and the indexical field. *Journal of Sociolinguistics 12,* 4, 453–476.

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.

NGUYEN, D., GRAVEL, R., TRIESCHNIGG, D., AND MEDER, T. 2013. "How old do you think I am?": A study of language and age in Twitter. In *Seventh International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*.

O'CONNOR, B., KRIEGER, M., AND AHN, D. 2010. TweetMotif: exploratory search and topic summarization for Twitter. In *ICWSM 2010*.

RAO, D., YAROWSKY, D., SHREEVATS, A., AND GUPTA, M. 2010. Classifying latent user attributes in Twitter. In *SMUC 2010*.

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.

Dong Nguyen is a PhD student at the University of Twente and affiliated with the Meertens Institute. She is interested in studying the relationship between language and society using computational methods and social media data.

Rilana Gravel studied English and German in Münster (Germany) and completed her Research Master in Linguistics at Leiden University in the Netherlands in 2013.

Dolf Trieschnigg is a postdoctoral researcher at the University of Twente and affiliated with the Meertens Institute. He is interested in various areas of information retrieval, information extraction and natural language processing. His PhD was on the integration of domain knowledge in information retrieval systems.

Theo Meder earned his PhD as a medievalist in 1991 at the University of Leiden. Since 1994, he works as a folklore researcher at the Meertens Institute in Amsterdam. He is specialized in folk narrative research and co-ordinator of the Dutch Folktale Database.