TweetGenie: Development, Evaluation, and Lessons Learned

Dong Nguyen[1] **Dolf Trieschnigg**[1] **Theo Meder**[2]
(1) Human Media Interaction, University of Twente, Enschede, The Netherlands
(2) Meertens Institute, Amsterdam, The Netherlands
{d.nguyen,d.trieschnigg}@utwente.nl, theo.meder@meertens.knaw.nl

Abstract

TweetGenie is an online demo that infers the gender and age of Twitter users based on their tweets. TweetGenie was able to attract thousands of visitors. We collected data by asking feedback from visitors and launching an online game. In this paper, we describe the development of TweetGenie and evaluate the demo based on the received feedback and manual annotation. We also reflect on practical lessons learned from launching a demo for the general public.

1 Introduction

The language use of speakers is related to variables such as the speaker's gender and age (Eckert, 1997; Eckert and McConnell-Ginet, 2013). Systems that can automatically predict such variables have been receiving increasing attention. They enable more fine-grained analyses of trends by profiling the involved users. They also support sociolinguistics research by shedding light on the link between variables such as gender and age, and the language use of speakers.

In this paper, we describe TweetGenie (www.tweetgenie.nl), a website that allows visitors to enter public Dutch Twitter accounts. The system predicts gender and age of the users behind the entered accounts based on the 200 most recent tweets. Due to press attention from various media outlets, we were able to attract a large number of visitors. In comparison to previous gender and age prediction systems that have been evaluated with carefully constructed datasets, we are the first to evaluate the performance of such a system 'in the wild'.

We first discuss the development of TweetGenie (Section 2). Next, we study the launch and TweetGenie's spread through social media, based on log data of the first week after the launch (Section 3). We then evaluate TweetGenie based on collected feedback (Section 4) and reflect on practical issues we encountered while launching an online demo for the general public (Section 5). We end with a conclusion (Section 6).

2 TweetGenie

In this section we describe the development and setup of TweetGenie.

Goals The original research (Nguyen et al., 2013) was carried out to support analyses of trends and to study sociolinguistic aspects of language use. By launching a public demo of this research, we aimed to 1) test the system on a large-scale 'in the wild' 2) collect data, and 3) demo the project to interested people. Unlike most demos of NLP research, the target audience of this demo was the '*general public*'. For example, we aimed for a simple and attractive interface, and released a press announcement to reach a large audience.

Model TweetGenie was developed based on the research and dataset described in (Nguyen et al., 2013) and predicts the gender and age of Dutch Twitter users based on the 200 most recent tweets. First, unigrams and bigrams are extracted from the tweets using the tokenization tool by O'Connor et al. (2010). This feature representation was chosen, because it is fast and unigrams have shown to perform

already very well (Nguyen et al., 2013). We then trained logistic (for gender prediction) and linear (for age prediction) regression models (Pedregosa et al., 2011) with L2 regularization.

Setup TweetGenie is available at www.tweetgenie.nl. After a visitor enters a public Twitter account, a results page is shown (see Figure 1 for a screenshot). The results page shows the predicted age (in years), the gender, and a gender 'score' indicating how strong the prediction was (based on $\mathbf{x}^\top \boldsymbol{\beta}$ with $\mathbf{x}$ being the features and $\boldsymbol{\beta}$ the estimated parameters). In addition, an option is available to share their results page on Twitter.

An overview of the components is shown in Figure 3. The first webserver hosts the frontend. A MySQL database is used to keep track of the progress of each prediction, and to store logs and feedback received by users. A second webserver is used to retrieve the data from Twitter and perform the predictions.

Feedback To collect data and improve the system, users are encouraged to provide feedback on the predictions. On the page with the automatic prediction (Figure 1), users have the option to enter the correct age and confirm whether the gender prediction was correct.

Online Game We also developed an online game to study how humans perform on the task. Figure 2 shows a screenshot of the interface. Players are shown 20-40 tweets per Twitter user and have to guess the gender and age of the Twitter users behind the tweets. After each guess, players receive feedback in various ways (the correct gender and age, the predictions by the automatic system, and the average guesses by other players). The data collected proved to be valuable: using the data, we reflected on the task of inferring gender and age from tweets and the limitations of current systems (Nguyen et al., 2014).

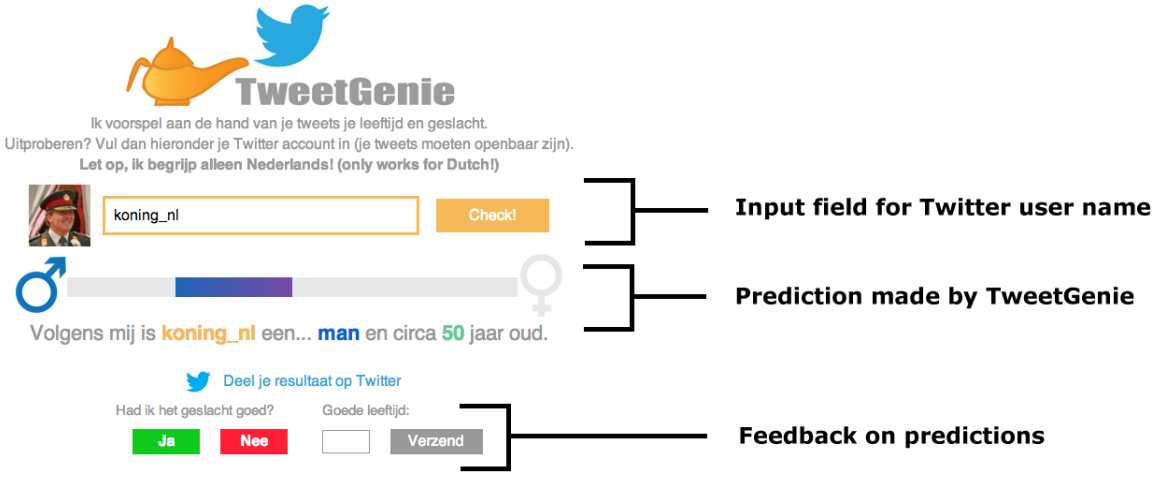


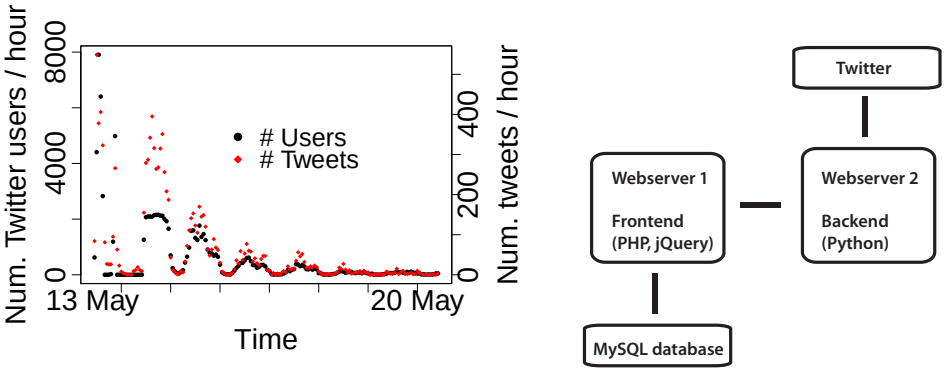
Figure 1: Screenshot prediction interface

Figure 2: Screenshot online game. Based on the shown tweets, players are asked to guess the gender and age of the user behind the tweets.

3 Launching TweetGenie

TweetGenie was launched on May 13, 2013 at around 11.30 AM. To reach a large audience, a press statement was released and messages were posted on social media networks. In this section, we analyze the data in the first week after the launch.

Figure 3 shows the number of entered Twitter users and the number of tweets mentioning TweetGenie in the first week after the launch. The number of tweets and the number of users entered follow similar trends. We observe a high peak in the beginning, but it also rapidly decreases over time. The system was asked to make a prediction 87,818 times and 9,291 tweets were posted with the word 'TweetGenie'. 1,931 of these tweets were created using the tweet sharing function of TweetGenie. The observed sentiment was mostly positive. If TweetGenie made an incorrect prediction, most people joked about it (e.g. '**grin* I just became 13 years younger without plastic chirurgy #tweetgenie*'). The game was played often as well, a guess was made 31,414 times.



(a) Entered users and tweets per hour (b) Architecture

Figure 3: Overview of the system

4 Evaluation

We evaluate the system in two ways, 1) using the feedback from users, and 2) using manual annotation.

4.1 Evaluation Based on User Feedback

Visitors were encouraged to give feedback on the predictions of TweetGenie. In the first week, we received feedback on the gender of 16,563 users and on the age of 17,034 users.

Reliability We randomly sampled 150 Twitter users for which we received feedback on both the gender and age. We checked the feedback of these users by visiting their Twitter profiles. If the feedback seemed plausible based on the profile, we assumed the feedback was correct (i.e. we did not visit any other social media profiles to find the exact age). The results are shown in Table 1. We find that 90% of the feedback appears to be correct. Only a small fraction (4%) of the feedback was incorrect, this could be deliberate or due to sloppiness. The remaining feedback was on Twitter accounts of non-Dutch users (e.g. English, German, French), or accounts that did not represent a person (e.g. a sports team, animal, multiple persons).

Accuracy We calculate the performance based on the 135 users for who we received correct feedback. We find that the users who gave feedback are *not* representative of the general Dutch Twitter population (Nguyen et al., 2013). The users are older than average (the age distribution is shown in Figure 4). There are more older males, and more younger females using Twitter in the Netherlands (Nguyen et al., 2013), and as a consequence the number of males (60.7%) is higher than the number of females (39.3%).

Based on this dataset, we find that the accuracy of the gender predictions was 94%. The Mean Absolute Error (MAE) for the age predictions is 6.1 years, which is higher than reported in (Nguyen et al., 2013).

Feedback	Frequency	Percentage
Correct	135	90%
Incorrect	6	4%
Not a Dutch account	5	3.33%
Not a person	4	2.67%

Table 1: Statistics feedback reliability

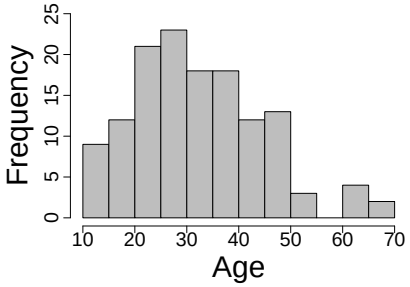


Figure 4: Age distribution feedback

However, this can be explained by the observation that relatively many older Twitter users give feedback, and as discussed in (Nguyen et al., 2013), automatic age predictions for older Twitter users are less accurate.

4.2 Evaluation Based on Manual Annotation

We also evaluated the system by manually annotating 50 users that were randomly sampled from the entered users in the logs. We did not include accounts that were not Dutch or did not represent individual persons. If feedback was available for a Twitter user, we used the provided feedback (after a manual check). Otherwise, we manually annotated the gender and age using all available information (e.g. social media profiles, websites). The gender was correctly predicted for 82% of the users, which is lower than measured in the evaluation based on the user feedback (Section 4.1). The Mean Absolute Error (MAE) is 6.18 years, which is in line with the observed MAE based on the user feedback.

Our analyses confirm that users for who feedback was available are *not* representative of all users who were entered in the system. Of the sampled 50 entered users, the fraction of males and females is almost equal (52% and 48%) compared to 60.7% and 30.9% in Section 4.1. The number of users who were less than 20 years old (15) is similar to the number of users in the range of > 20 and ≤ 30 years (17), while in Section 4.1 the fraction of users below 20 years is smaller. Thus, less feedback was received for younger Twitter users.

In line with the analysis in Section 4.1, we find that relatively many older Twitter users were entered into TweetGenie compared to a more representative set of Dutch Twitter users (Nguyen et al., 2013).

5 Lessons Learned

We learned many lessons from launching a demo for the general public.

1) *Test all components of the demo*. While developing the system, we focused mostly on ensuring that the backend would be able to handle the number of visitors. However, after the demo went online, problems arose at the frontend due to the visitor load. This was solved by only allowing a fixed number of visitors at the same time. We also did not test the interface for non-Dutch visitors. Only later we found out that the automatically translated version contained serious errors: international visitors were misled that the model worked on English tweets.

2) *The distribution of users trying out the demo might not correspond to the distribution in the development dataset.* While we extensively evaluated the system on a carefully constructed, representative dataset (Nguyen et al., 2013), the numbers in this paper's evaluation are lower. Users who were entered into the system were not representative of the Dutch Twitter population: relatively more older Twitter users were entered in the system, leading to more errors in the automic age prediction.

3) *A demo is a good opportunity to collect data.* Many visitors were willing to provide feedback or participated in the online game. Data collected through the online game has been used to study the task of inferring gender and age in more depth (Nguyen et al., 2014). Manual analysis of the feedback in this paper revealed that almost all of the feedback appears to be genuine. Further research is needed to study how the feedback on the automatic predictions can be used to improve the prediction models.

6 Conclusion

In this paper we discussed TweetGenie, an online system that infers the gender and age of Twitter users based on tweets alone. We collected much feedback from the users, but also found that users who provided feedback are not representative of all the entered users. We demonstrated that besides being a valuable tool for user profiling, TweetGenie also appeals to the general public.

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).

References

P. Eckert and S. McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

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

D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. 2013. "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of ICWSM 2013*.

D. Nguyen, D. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F.M.G. de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014*.

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

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