Analyzing crowdsourced assessment of user traits through Twitter posts

Lucie Flekova[*]
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt
flekova@ukp.informatik.tu-darmstadt.de

Salvatore Giorgi
Jordan Carpenter
Positive Psychology Center
University of Pennsylvania

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

Abstract

Social media allows any user to express themselves to the public through posting content. Using a crowdsourcing experiment, we aim to quantify and analyze which human attributes lead to better perceptions of the true identity of others. Using tweet content from a set of users with known age and gender information, we ask workers to rate their perception of these traits and we analyze those results in relation to the crowdsourcing workers' age and gender. Results show that female workers are both more confident and more accurate at reporting gender, and workers in their thirties were most accurate but least confident for rating age. Our study is a first step in identifying the worker traits which contribute to a better understanding of others through their posted text content. Our findings help to identify the types of workers best suited for certain tasks.

Motivation

Large scale user generated content allows us to study language use in a richer context than ever before, including the attributes of a user such as demographics, personality and spatio-temporal information. By assuming language use is influenced by user attributes, previous research used posts from a user to build statistical models to infer different characteristics such as: age (Rao et al. 2010), gender (Burger et al. 2011) or occupation (Preoţiuc-Pietro, Lampos, and Aletras 2015). Applications of automatically inferring user traits range from recommender to dialogue systems which can produce tailored content to different user groups. To train these models, researchers used gold labels either extracted from user self-reports (Preoţiuc-Pietro, Lampos, and Aletras 2015) or predicted by workers (Volkova et al. 2015). However, researchers acknowledge that there is variation both in the expression of trait cues in authors' text as well as readers' skill in interpreting these cues (Kenny and Albright 1987), leading to statistical methods that model either the real attributes or their perception.

In this study, we present a crowdsourcing experiment on annotating user traits from Twitter posts. Rather than focusing on predictive performance, we analyze the crowdsourcing workers' own attributes and how these influence prediction performance and confidence. For example, do certain worker traits lead to better predictions and higher confidence, or which users are overall easiest to identify?

We use age and gender as the target traits for both annotation and worker analysis, as they are considered basic categories in person assessment (Quinn and Macrae 2005) and are highly studied by previous research. (Nguyen et al. 2014) studied the crowdsourcing performance of predicting age and gender, highlighting that this is a hard problem once the target age is over 30, and that teenage users are estimated to be older. At the worker level, (Kazai, Kamps, and Milic-Frayling 2012) analyzed multiple user traits in the context of crowdsourcing relevance labels, finding that geography plays a major role in performance with smaller effects for gender and age. In general, user demographics and behaviour predict which users are more trustworthy (Kazai, Kamps, and Milic-Frayling 2011), while (Li, Zhao, and Fuxman 2014) proposes a method to target tasks to specific user groups.

Methodology

We study the worker prediction of two user traits – gender and age – through Twitter posts. In the annotation task, we use a set of posts from users previously matched to their true age and gender. For gender, we use the users from (Burger et al. 2011), which are mapped to their self-identified gender by linking them to their other public profiles. This dataset consists of $67,337$ users, from which we create a balanced sample of 1000 users. The age dataset is obtained by identifying $4,279$ users that were the target of a tweet such as 'Happy X birthday to @USERNAME'. For our experiment, we sampled 1000 users from each of the five culturally meaningful age groups: < 18, $18 - 22$, $23 - 30$, $31 - 40$, $41+$.

We created an annotation task on Amazon Mechanical Turk. Each HIT consisted of 20 tweets sampled from a pool of 100 tweets posted by each user over the past six months. The workers predicted either age or gender, stating the confidence of their rating on a scale from 1 to 5. Each user was assessed independently by 9 workers. We administered a questionnaire to collect worker information. For quality control, we used a set of HITs where the age or gender was explicitly stated within the top 10 tweets displayed to the worker. The control HIT appeared 10% of the time and a worker missing the cor-

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

rect answer twice was excluded from annotation and all his HITs invalidated. Further, we limited the location of workers to the US. An example HIT is presented in Figure 1 at `http://bit.ly/1LFpDx8`. In total, we obtained 38.7% HITs from male workers on gender and 14.6% from 18-22, 49.8% from 23-30, 25.8% from 31-40 and 9.6% from 40+ year old workers on age.

Analysis

We first analyze the performance of gender prediction across worker's genders. Table 2 shows the gender predictions at HIT level, separated out by worker gender. We can conclude that females are better at predicting gender overall. Analyzing the errors, we see that males have lower performance mostly due to failing to make accurate predictions when rating males. Overall, females are easier to accurately rate than males (.392/.403 vs. .339/.347). Additionally, females have a higher overall self-reported confidence in their prediction, even when the prediction is incorrect. The highest relative increase in confidence is when females predict other females (3.85 vs. 3.65 for male workers), which is where the highest decrease in error is also observed (.138 vs. .159) compared to their male counterparts.

Real/Pr.	**Male**	**Female**	C_M	C_F
Male	.339 / .347	.159 / .138	3.26 / 3.37	3.18 / 3.31
Female	.110 / .112	.392 / .403	2.97 / 3.06	3.65 / 3.85

Table 1: Left part of the table displays normalized confusion matrices of workers prediction of gender. Right part of the table displays average self-reported confidence on those prediction groups. In both cases, the values in a cell show the performance of male (left) and female (right) workers respectively.

In terms of age, the best precision is reached for the raters in the $31 - 40$ class, followed by the $23 - 30$ class. We also notice the trend across all workers, regardless of their own ages, have similar prior beliefs about the user distribution on Twitter, while adjusting towards their own age group when unsure (indicated by higher recall and lower precision). The easiest class to predict are users between $23 - 30$ years old. Intriguingly, based on the self-identified user confidence in the ratings, users between $31 - 40$ are the least confident (3.20), compared to users aged $23 - 30$ (3.26) who were second best at prediction and second least confident. The groups $18 - 22$ (3.38) and $40+$ (3.46) were most confident and least accurate.

Pred./worker	<18	18-22	23-30	31-40	41+
18-22	.312 / .221	.555 / .336	.369 / .441	.163 / .366	.081/ .166
23-30	.345 / .219	.491 / .346	.353 / .434	.177 / .311	.172 / .217
31-40	.273 / .266	.518 / .359	.402 / .443	.201 / .336	.271 / .326
41+	.202 / .229	.463 / .394	.384 /.436	.178 / .194	.269 / .194

Table 2: Performance of the workers of each age class (row) on each twitter user's age class (column). First value in a cell displays recall, second value precision. Note that precision decreases and recall increases as workers approach the class of the user.

Conclusions and Future Work

We presented an analysis on the impact of worker age and gender in the task of user trait prediction from Twitter posts. Demographic differences exist, leading women to be more accurate and confident overall in gender prediction. However, this is determined primarily by females being more accurate when identifying other females. For age, users in their thirties are both more accurate and less confident than all others. These results suggest worker populations that are most successful at different user categorization tasks.

We aim to continue our study by crowdsourcing prediction for other user traits – specifically education level and political orientation. In our worker qualification, we collected further demographic information including education level as well as psychological questionnaires. These can be used to gain further insight into the psychological traits which influence the workers' ability to make better guesses.

Another avenue for future research is to automatically quantify the linguistic features that lead to worker's ratings. We aim to compare these features to those of machine learning models built on gold standard labels in order to identify the linguistic markers that mislead workers' ratings. This would allow to gain a better understanding of text use and to develop tailored content for different demographic groups, which is very important for recommender systems, dialogue agents or tailored education systems.

Acknowledgements

The authors acknowledge the support from Templeton Religion Trust, grant TRT-0048.

References

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

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2011. Worker types and personality traits in crowdsourcing relevance labels. In *CIKM*.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *CIKM*.

Kenny, D. A., and Albright, L. 1987. Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*.

Li, H.; Zhao, B.; and Fuxman, A. 2014. The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *WWW*.

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

Preoţiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *ACL*.

Quinn, K., and Macrae, N. 2005. Categorizing others: The dynamics of person construal. *Journal of Personality and Social Psychology*.

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

Volkova, S.; Bachrach, Y.; Armstrong, M.; and Sharma, V. 2015. Inferring latent user properties from texts published in social media (demo). In *AAAI*.