Detecting Demographic Bias in Automatically Generated Personas

Joni Salminen
Qatar Computing Research Institute, Hamad Bin Khalifa University; and Turku School of Economics
Doha, Qatar
jsalminen@hbku.edu.qa

Soon-gyo Jung
Qatar Computing Research Institute, Hamad Bin Khalifa University
Doha, Qatar
sjung@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University
Doha, Qatar
bjansen@hbku.edu.qa

ABSTRACT

We investigate the existence of demographic bias in automatically generated personas by producing personas from YouTube Analytics data. Despite the intended objectivity of the methodology, we find elements of bias in the data-driven personas. The bias is highest when doing an exact match comparison, and the bias decreases when comparing at age or gender level. The bias also decreases when increasing the number of generated personas. For example, the smaller number of personas resulted in underrepresentation of female personas. This suggests that a higher number of personas gives a more balanced representation of the user population and a smaller number increases biases. Researchers and practitioners developing data-driven personas should consider the possibility of algorithmic bias, even unintentional, in their personas by comparing the personas against the underlying raw data.

CHI'19 Extended Abstracts, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5971-9/19/05.
DOI: https://doi.org/10.1145/3290607.3313034

KEYWORDS

Automatic Persona Generation; Data-Driven Personas; Algorithmic Bias

Bias in Automatically Generated Personas – The Main Question
Data-driven personas area meant to correct for human biases, but are they in fact (a) introducing more biases or (b) reinforcing the existing biases? While these questions are difficult to answer in the scope of one study, we focus here on one aspect of data-driven personas and biases, namely: *Is there a bias in automatically generated personas?*

Defining Bias in Data-Driven Personas
In our context, we define bias as not showing some demographic groups in the automatically generated persona profiles, even though these demographic groups *would be* relevant to end users. We define "would be" such that the demographic groups are more prevalent in the eligible groups than in the representative groups chosen for the generated personas.

1 INTRODUCTION

"Data-driven personas" have the connotation of being plausible, credible, and trustworthy. Traditionally, it has been thought that data-driven, quantitative approaches, and techniques are more objective and give more valid results than qualitative or subjective methods [3]. However, recently this naïve belief in algorithmic decision making has been challenged, as researchers have become aware of algorithmic bias and the societal risks of machine decision making [4].

Essentially, the challenge of bias in algorithmic decision-making extends across all fields and applications using a substantial degree of automation in the place of human judgment. Examples of research focused on algorithmic bias include prevention of racial bias in the judicial system, fair credit scoring of consumer loans, and equal opportunity to education and job admissions.

A particular risk of bias exists when automatically creating data-driven personas. An example of such development is the automatic persona generation (APG) methodology introduced by An et al. [1,2]. The said methodology automates the persona generation process, removing manual labor in data collection, analysis, and persona creation. The methodology promises faster, more accurate, and updatable personas by eliminating the "tedious" manual data collection and analysis steps that are time-consuming and susceptible to human bias. However, while APG's benefits have been argued for in prior research [1,2], researchers have not previously investigated the existence of bias in these automatically generated personas.

In this research, we automatically generate different numbers of personas from a dataset collected from a large YouTube channel, those personas representing the content consumption patterns of various demographic groups. We then analyze if any demographic groups are absent in the generated personas, report the findings, and finish by discussing the implications of bias in data-driven personas as well as suggesting means for preventing it.

Demographic bias, in our context, refers to (a) age, (b) gender, and/or (c) country under- or overrepresentation in the generated personas, and our research question is: *Are some demographic groups over- or underrepresented in automatically generated personas?*

2 RELATED LITERATURE

The history of personas is closely associated with the notions of bias and stereotyping. More particularly, we can distinguish three perspectives to bias in personas from earlier literature: (a) the concept of persona being biased by definition; (b) persona are creators biasing the personas through stereotyping; and (c) the end users of personas are transferring their own biases and stereotypes to the created personas [8].

The choice for persona creators then, as noted by Chapman and Milham [3], is to make trade-off decisions – most typically, choosing the most dominant attributes of the user base instead of focusing in its full diversity. The choices by persona creators can be biased by their intentional or unintentional bias of selecting information, stereotypical thinking arising from personal experiences and attitudes, or personal ambitions within societal power structures [6]. Persona creation and usage can also include political motives and agendas, especially within organizational contexts [7].

Steps of Automatic Persona Generation

1. Creating an interaction matrix with videos as columns and demographic groups as rows

2. Applying NMF to interaction matrix to discern latent video viewing patterns

3. Choosing the representative demographic group for each pattern by using NMF weights

4. Creating the personas by enriching the representative demographic groups with additional information (e.g., name, picture, topics of interest).

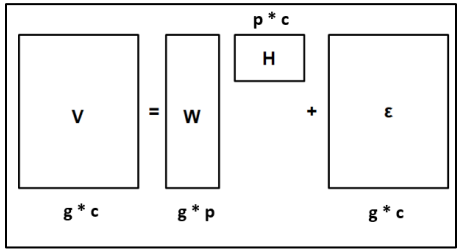

Figure 1: Matrix decomposition carried out using NMF. Matrix V is decomposed into W and H. *g* denotes demographic groups (respondents in the dataset), *c* denotes videos, and *p* is the number of latent video viewing patterns (i.e., skeleton personas).

As evident from prior persona studies, the reason for bias in personas has often been attributed to their human creators. To remedy the role of humans, "objective", data-driven techniques have been proposed by several scholars (see review in [2]). However, given the evidence of algorithmic bias in other contexts of automation, it is not certain whether automation actually mitigates bias in persona creation. This is because while automation can be used to mitigate human bias, it can introduce another form of bias, i.e., algorithmic bias. Because the previous literature employing data-driven personas does not explore the possibility of bias in the outcome personas, investigating the existence of bias in data-driven personas represents an important research gap.

3 METHOD

3.1 Data Collection

To investigate the potential of bias in automatically generated personas, we collect an example dataset of real user interactions with online content and use this dataset to automatically generate personas using the APG approach of An et al. [1,2]. This approach is chosen as it represents the state-of-the-art in data-driven persona development.

The data for persona generation was collected from the YouTube channel of Al Jazeera Media Network (AJ+) using the YouTube Analytics API. The dataset includes all the channel's view counts of 13,251 videos published between January 1, 2016 and September 30, 2018. The data shows 206,591,656 video views. The view counts are broken by demographic groups (age group x gender x country). There are 1,631 demographic groups with at least 1 view count out of 3,486 maximum combinations provided by the YouTube Analytics platform. Therefore, we can conclude that the data represents a diverse audience.

Note that while the data collected for this research is from the YouTube platform, the methodology of APG can be applied to any data structure that presents an interaction metric by a group or individual users over the content or other items (e.g., products).

3.2 Persona Generation

We use the collected YouTube data to generate personas that describe the channel's audience. We apply the method by An et al. [2] that allows the creation of an arbitrary number of personas using non-negative matrix factorization (NMF) (see Figure 1). Persona generation steps are summarized in the sidebar.

The personas are generated by first transforming the collected YouTube data into an interaction matrix, where rows of the matrix represent demographic groups (age x gender x country) and columns represent the content pieces, in this case, each individual video of the channel. The elements of the interaction matrix are video view counts. For example, reading the matrix, one could tell that "Men, aged 25-34, United States" have watched "Video X" in total 2,345 times as a group. An example of the data structure is provided in Table 1.

Table 1: Example of interaction matrix structure. Content is in columns and demographic groups are in rows. Cells denote how many times a video has been watched by a group.

Group	*Video 1*	*Video 2*	*Video n*
Male, 25-34, US	1,501	2,800	300
Female, 18-24, PH	3,333	1,700	10,250

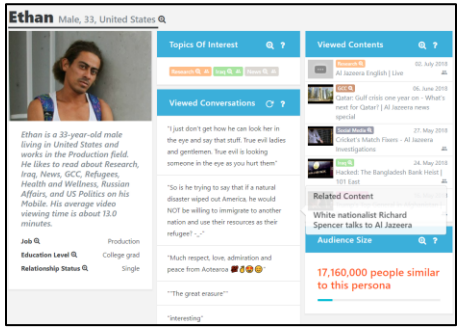

Figure 2: Example persona ("Ethan") obtained via APG. The information corresponds to typical persona profile, containing e.g. picture, name, age, descriptive quotes, and topics of interest.

Gender Bias Calculation

$$Gender\_bias_{i,j} = \frac{Gender\_personas_{i,j}}{Total\_personas_i} - \frac{Gender\_weights_{i,j}}{Total\_weights_i}$$

i = P20, P50, P100, P150, P200, P250, P300
j = Male, Female
$Gender_{weights}$: sum of NMF weights.
$Gender\_{personas}$: number of representative personas

For gender comparison, we take the sum of the NMF weights by gender and persona set and calculate the share of each gender from the total sum. Then, we calculate the share of each gender in the representative persona table and work out the difference between the two to arrive a bias for each gender.

The APG algorithm then utilizes NMF to discern p latent patterns from the interaction matrix, and outputs the p patterns as "skeleton personas," where each persona has demographic groups sorted by weight value that describes the strength of association between the demographic group and the latent pattern (i.e., skeleton persona). Each persona has its representative demographic group, selected among possible demographic groups. The APG algorithm seeks the representative demographic group for each persona sequentially from 1 to p, where the chosen demographic group is the one with highest NMF weight. Note that if the algorithm finds a duplicated demographic group for the group of p personas, it chooses the demographic group with the second highest weight value. An example of the APG personas is provided in Figure 2.

3.3 Bias detection in automatically generated personas

Intuition: In the following, we present a simple approach for detecting bias in automatically generated personas, and validate it using the collected dataset. The intuition is that the demographic groups ranking the highest by their NMF weights should be chosen as representative groups when generating a set of personas.

Consider that a set P consists of 1 ... p personas. Each $p \in P$ consists of 1...g demographic groups, where the maximum of g is defined by the product of [age x gender x country], i.e., the demographic attributes.

Each set of personas thus contains $G = \sum_{n=1}^{p} g$ groups, each having an associated NMF weight value. Within a persona set, we can thus generate p personas and take g ∈ G groups from each persona by sorting their NMF factorial weights from highest to lowest.

This results in a set of G groups. For example, if each persona contains five demographic groups and there are three personas, then $G = 5 \times 3 = 15$. We can see that $G > P$, which is the root cause for potential bias since the representative demographic group for a persona is chosen from G.

Bias detection: To detect potential biases, we average the NMF weight values of the G groups across p personas, and sort the G groups from highest to lowest by their overall weight value that we denote by φ. Using φ, we take Top N groups from the G groups, i.e., the N overall highest-ranking groups. Here, N is set equal to p, i.e., $N = p$.

Then, we examine how many of these groups are actually selected as representative persona demographic groups. For example, if a Group 1 would be included in the Top N group, but not belong to the representative persona group, then that group is "discriminated" against (because it has a globally high NMF weight within the persona set but is not chosen as the demographic group of any of the personas in the set).

4 RESULTS

To query the potential bias in APG personas, we use the dataset to generate seven persona sets: P_{20}, P_{50}, P_{100}, P_{150}, P_{200}, P_{250}, and P_{300}, where the index number indicates the number of personas in the set. Again, we remind that the selection of demographic attributes is done automatically by the system, based on the inputted data. Figure 3 illustrates the selection.

Male 25-34 United States	1080.30
Male 35-44 United States	350.15
Male 25-34 United Kingdom	341.00
Male 25-34 Canada	324.75
Male 25-34 India	243.31
Male 25-34 Australia	195.61
Male 18-24 United States	112.33
Male 25-34 Germany	100.84
Female 25-34 United States	86.56
Male 35-44 Canada	77.57

Figure 3: Top 10 demographic groups of "Ethan". The values are NMF weights that indicate how well the demographic group corresponds with the inferred latent video viewing pattern. The demographic group with the highest value is chosen as the representative demographic group for the persona.

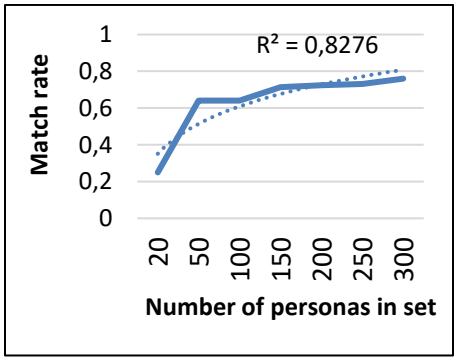

Figure 4: Varying the number of personas results in a logarithmically increasing function. In other words, the match rate between Top *N* personas and representative personas increases with the number of personas in a set.

First, we compute the simple match as $MatchRate = R / P$, where R is the number of demographic groups that are both in Top N and representative group, and P is the number of personas (and, as such, the number of representative groups chosen). In other words, if all the personas in the Top N are found in the representative group, the match rate is 1.0 (100%). If no Top N groups are chosen as representative groups within a persona set, then the match rate is 0.0 (0%). Figure 4 shows the match rates across the persona sets. Overall, the match rate increase with the number of personas. In P_{20}, the match rate is only 25.0%, meaning that most highest-ranking demographic groups are not chosen as representative groups. In turn, for P_{300}, the match rate is 76.0%, indicating that most highest-ranking groups are chosen.

Next, we examine the age correspondence. We do this by first aggregating the numbers. For example, two personas, male 25-34 and female 25-34, would no belong to the same age group, i.e. 25-34. We, therefore, compare the age group match between the representative and Top N groups. We use the Pearson correlation coefficient to compare the counts of instances from each group (see Table 2 for example). The results can be seen in Table 3.

Table 2: Example of count data used for comparing correspondence of Top N and demographic groups' ages. The example is from P_{20} persona set.

Age group	*Top N*	*Representative*
13-17	0	0
18-24	7	4
25-34	11	14
35-44	1	2
45-54	1	0
55-64	0	0
65-	0	0
total	20	20

Table 3: Pearson correlation coefficients for Top N and representative demographic groups

	P_{20}	P_{50}	P_{100}	P_{150}	P_{200}	P_{250}	P_{300}
Pearson correlation	0.94	0.93	0.96	0.99	0.98	0.99	0.99

Regarding age, we also observe a general tendency that age diversity increases with the number of personas (see Figure 6). For example, age groups 13-17, 45-54, 55-64, and 65-> are completely missing from the P_{20} persona set. Even though the dominance of age group 25-34 is consistent throughout the persona sets, the color coding in Figure 6 indicates that more age groups are shown when the number of personas increases.

Finally, we compare the gender representation in Top N and representative groups. For this comparison, we take the sum of the NMF weights by gender and persona set and calculate the share of each gender from the total sum. Then, we calculate the share of each gender in the representative group, and we deduct this number to the previous number. (See the sidebar “Gender Bias Calculation” in the previous page for formal definition.)

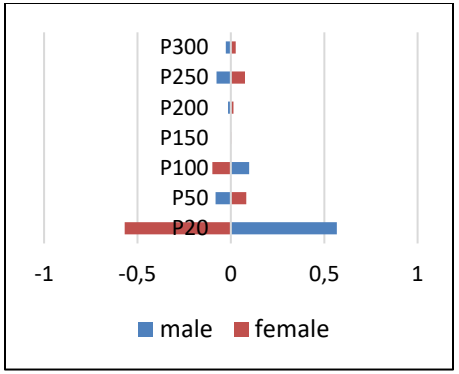

Figure 5: Under- or overrepresentation of each gender by persona set. The bars indicate deviation from zero, the ideal situation where the match of genders between the representative and Top N groups is perfect.

Takeaways and Future Work

The results show a relationship between the number of personas generated and the demographic bias. This was visible in all levels of testing: exact match, age, and gender.

The results raise the question of how many personas should be generated in order to better capture the diversity of user segments in the data. The fewer personas are generated, the more room there is for bias. However, increasing the number of generated personas, while truer to data, might be confusing for end users, as many personas are difficult to remember and work with. Future research should investigate this trade-off, both from technical point of view as well as from persona user perspective.

Future work could also present a more thorough comparison of the techniques applied here with other possible methods. Finally, a formative study with persona users would be beneficial. For example, observing how they might interact with an underrepresented group could lead to additional insight on bias in the persona context.

Age Group	P20	P50	P100	P150	P200	P250	P300
13-17	0	1	4	6	6	10	9
18-24	4	8	19	31	43	55	60
25-34	14	29	51	70	88	100	119
35-44	2	5	14	26	34	47	61
45-54	0	3	4	6	12	18	25
55-64	0	3	6	8	10	13	14
65->	0	1	2	3	7	7	12

Figure 6: Counts of age groups through the persona sets

As visible from Figure 5, APG is fairly robust against gender bias, as the gender distributions of Top *N* and representative groups are very similar. The exception is P_{20}, where only two personas are female, but the NMF weights are in fact higher for women. This is because there are relatively few female demographic groups compared to male, but the ones that are present have higher weights than average male demographic groups.

REFERENCES

[1] Jisun An, Haewoon Kwak, Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen. 2018. Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. Social Network Analysis and Mining 8, 1.

[2] Jisun An, Haewoon Kwak, Joni Salminen, Soon-gyo Jung, and Bernard J. Jansen. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. ACM Transactions on the Web (TWEB) 12, 3.

[3] Christopher N. Chapman and Russell P. Milham. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 50, 5: 634–636.

[4] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, 432.

[5] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2125–2126.

[6] Nicola Marsden and Maren Haag. 2016. Stereotypes and Politics: Reflections on Personas. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, ACM, 4017–4031.

[7] Kari Rönkkö. 2005. An Empirical Study Demonstrating How Different Design Constraints, Project Organization and Contexts Limited the Utility of Personas. Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences - Volume 08, IEEE Computer Society.

[8] Phil Turner and Susan Turner. 2011. Is stereotyping inevitable when designing with personas? Design studies 32, 1: 30–44.

[9] Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values 41, 1: 118–132.