All About the Name: Assigning Demographically Appropriate Names to Data-Driven Entities

Soon-gyo Jung
Qatar Computing Research Institute
sjung@hbku.edu.qa

Joni Salminen
Qatar Computing Research Institute; and Turku School of Economics
jsalminen@hbku.edu.qa

Bernard J. Jansen
Qatar Computing Research Institute
bjansen@hbku.edu.qa

Abstract

We develop a method for assigning demographically appropriate names to data-driven entities, such as personas, chatbots, and virtual agents. The value of this method is removing the time-consuming human effort in this task. To demonstrate our method, we collect four million user profiles with gender, age, and country information from an international online social network. From this dataset, we obtain $1,031,667$ *unique names covering* $3,088$ *demographic group combinations that our method considers as gender, age, and nationality appropriate. A manual evaluation by raters from 34 countries shows a demographic appropriateness score of 85.6%. The demographically appropriate names can be utilized for data-driven personas, virtual agents, chatbots, and other humanized entities.*

1. Introduction

Researchers working on computational social science and user modeling use demographic information, such as age, gender, and country, to better understand user behavior [1, 2, 3]. When demographic information is not readily available, it needs to be inferred from other information, such as the user's name [4]. This is defined as the demographic inference problem, focused on discovering data sources and developing algorithms towards the reliable prediction of user demographics from secondary data.

However, there are cases when the reverse takes place: the demographics of the user (or user segment) are known, and we need a name to represent those users. In particular, demographically appropriate names are needed for enhancing the realism of virtual entities in systems and software applications [5, 6]. These entities rely on anthropomorphism to enhance end-users' immersion with those systems [7]. For example, social media audiences can be represented as data-driven user segments – defined as groupings of users based on shared behaviors or characteristics [8]. One user segment type is data-driven personas (DDPs) whose purpose is to communicate numerical analytics data in a personified way to human decision-makers – essentially, giving faces to data [9].

A DDP is a fictive person that describes a user or social media segment [10, 11] through a 'bio' type profile containing various information, including picture, name, age, gender, and nationality of the persona. DDPs are used in domains, such as software development, user experience, design, and marketing. DDPs make it possible for decision makers to see use cases 'through the eyes of the users' and facilitate communication among team members through shared mental models [12].

Manual creation of user segment such as DDPs is a time-consuming and lengthy process [13]. To mitigate the need for manual labor, researchers have developed methodologies for development of DDPs. These methodologies generate DDPs from social media and other online data and present the personas to decision makers via an online user interface. Automatic Persona Generation (APG) is one of the data-driven methodologies [13, 14, 5]. Figure 1 illustrates a DDP created using APG.

This increasing popularity of DDPs [15, 16] results in a critical and challenging need for automatically assigning names that are demographically appropriate for each generated DDP, as the name is one of the key attributes increasing the perceived authenticity of a persona for decision makers [17]. The name is important because it affects the perceived realism by the users viewing the humanized entities. This is particularly important in the case of personas, as the name (along with a profile picture) form the basis of the typical persona profile [18]. The inappropriate choice of the name can increase in the perceived inconsistency of the persona profile, as the name would not match the demographic information of the persona.

However, assigning demographically appropriate names to DDPs, or similar entities requiring realistic names (e.g., virtual agents, humanized chatbots),

URI: https://hdl.handle.net/10125/71108
978-0-9981331-4-0

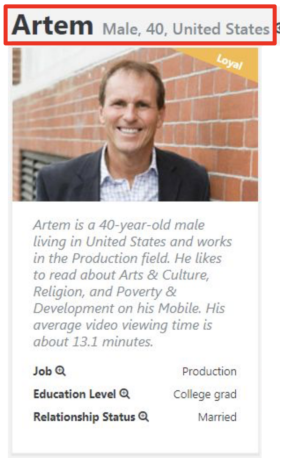

Figure 1. DDP profile automatically generated by APG: Is 'Artem' the authentic name for the demographic information? When automatically creating human-like entities such as personas, virtual agents, and chatbots, one needs to ensure names correspond to the demographic features of the entity.

is difficult, as manual verification of demographic appropriateness is costly at scale. Nevertheless, a name is a critical part of these entities – it gives an impression of a "real person" that is instrumental for empathizing with and relating to the user segment (or chatbot, virtual agent) at a personal level [12].

The problem we tackle is: ***How to identify demographically appropriate names for artificial entities using social media data?***

While the name identification problem is applicable to different artificial entities, such as virtual agents and chatbots [19, 20], we address the problem in the context of user segmentation. User segmentation, with the help of data-driven entities such as DDPs, is applied in many domains, including marketing and advertising, sales, and public health [15]. In these domains, online audiences can be represented as data-driven user segments. These data-driven segments, in turn, can be presented as personas, with a name and demographics attributes. For this, the names need to be demographically appropriate, so that the user segments appear authentic. Major social media actors, such as news and media channels, have diverse audiences, which requires a large number of names – more than can be manually curated.

Adopting the use case of DDPs, we demonstrate a methodology that addresses this problem for personas and similar entities requiring a human name.

2. Related Work

Researchers have developed various methods for automatic detection of users' demographics [4, 21] on social media to understand the online populations and behaviors [22, 23]. One way to infer demographics is a utilizing names of social media users [22, 24, 25]. The limitation of the methods is that insufficient baseline data for the prediction may result in less precise prediction at scale [22]. For demographically diverse audiences – such as many age groups in many countries – this results in a compromise between sufficient observations within different demographic groups and the precision of the demographic predictions.

Liu and Ruths [26] found that using the first name of Twitter profiles when predicting the user's gender improved the accuracy by 20% relative to a standard baseline classifier. Cunha et al. [27] inferred users' gender by comparing the first names provided in the Twitter profiles of users with a list of gender-appropriate names. Their method achieved an 89.1% accuracy. However, the approach was focused on first names in Portuguese, and thus cannot be readily applied to large international audiences.

Mueller and Stumme [28] proposed a model called NamChar that uses phonetic features of the name to predict the gender of a user. Their approach achieved a 69.2% accuracy on the test set. Huang et al. [6] trained a machine learning model to predict a user's nationality. One of the used features was name-ethnicity (NE), a 10 dimensional vector with each dimension representing an ethnicity. When training the classifier, NE had the highest rank of normalized feature importance among the tested 17 feature types. The researchers interpret the findings such that NE provides a useful signal for inferring a user's nationality in most cases.

Oktay et al. [29] estimates age and ethnicity of users based on their first names. Their exploratory results suggest that a name is a useful indicator for this purpose, although the results are limited to data from the United States. Hofstra and Schipper [22] uses Facebook data predict the most likely ethnicity value from first names, while quantifying the uncertainty of the prediction.

Collectively, results of the previous work suggest that it is hard to obtain all the realistic names for the numerous and diverse audiences of major social media platforms. We propose this as first research gap (**Gap 1**) that our research addresses.

Furthermore, while previous work has made progress with inferring *one demographic trait* – age, gender, or nationality – from social media profiles and user-generated texts at a time, there is no a high-performing solution for inferring *all three*

attributes at once. This is the second research gap (**Gap 2**) that our research addresses. A third research gap (**Gap 3**) is the limited geographic focus – for example, focusing only users from the United States [29]. To represent large and diverse international audiences, such as in creating DDPs from the YouTube followers of a major news and media channel [30], coverage of different nationalities is required.

Another limitation of the methods, especially when applying algorithms, is the lack of ground truth data for evaluation. The names and their associated demographics might not be readily available for every country in test sets [22]. To overcome this, researchers have applied Census data for testing [28, 29]. However, Census data on names is not available for all countries, and it is typically not coupled with both gender and age attributes. Another alternative researchers have used is manually annotating the data with crowdsourcing [31]. However, large-scale crowdsourcing can be costly, fallible, and the crowdsourcing platforms may not have crowd workers from all the necessary geographic regions to provide high-quality labels. Thus, the lack of evaluation datasets is the final research gap (**Gap 4**).

In combination, Gaps 1-4 suggest that there is no available solution for the practically important (and theoretically interesting) problem of demographic attribute inference of user profile names. By using a large, demographically diverse dataset of real names with associated demographics, we develop an inference approach that addresses these research gaps.

The research purpose is to identify realistic names for a large variety of demographic combinations by using information from online social media profiles. Consequently, we aim to make it possible to automatically assign names to DDPs and other user segments or entities with demographic characteristics. For this, we provide an easy-to-use online tool – **GAN2Name**[1] – for generating demographically appropriate names.

3. Methodology

3.1. Data Collection

We use InterPals[2] to collect names and corresponding demographic attributes. InterPals is a social network for language exchange among native speakers. InterPals asks all users for name, birthday, gender, and hometown and current city/country information. The basic demographic information is displayed on the user's profile page.

The names and their demographic attributes are retrieved using Scrapy, an open-source web-crawling framework for Python. Using this tool, we visit the profile pages and extract name, age, gender, hometown, and current country code (ISO 3166). In total, we collect $3,817,272$ profiles with demographic attributes. Table 1 shows some examples. Apart from the name, no personal information, such as email, phone number, or pictures was collected. Also, we only use the first names of the users to mitigate any harms from using personally identifiable information [32].

Name	Age	Gender	Hometown	Current
Ploy	29	female	TH	DE
Shady_Decker	15	male	-	US
Kim	37	male	-	KR

Table 1. Sample profiles with demographic attributes collected from InterPals

3.2. Data Preparation

Here, we present the process of cleaning and transforming the collected profiles data into a ready-to-use dataset which can be used for the implementation of our methodology with four demographically measurable indicators.

First, we apply a list of age bins used by most data analytics platforms (e.g., Google Analytics, YouTube Analytics) and the APG application; the age bins are (13-17), (18-24), (25-34), (35-44), (45-54), (55-64), and (65+). Sixty-three profiles are removed from the data as their ages are below 13. We choose a country code for each profile based on hometown or current country code. If a profile has a hometown country code, the profile takes the hometown country code as its country code. There are $52,741$ profiles (1.38% among entire profiles) that take the hometown country code as their country code. For instance, the profile for 'Ploy' takes 25-34 as its age bin and TH as its country code in Table 1.

Second, numeric characters and abbreviated names (e.g., J. W.) are removed. The special character period '.' and underscore '_' are replaced with white space considering as a link of two different words. 'Shady_Decker' is one of examples, so it becomes 'Shady Decker' in Table 1. Special characters are removed, except hyphen '-', which can be a part of the given name in some countries [3].

Third, we filter out the family names from the collected InterPals profiles. For this purpose, we use SPARQL of Wikidata[4] to collect $231,726$ data items consisting of family name and its corresponding country

[1] https://quecst.qcri.org/tool/GAN2Name
[2] https://interpals.net
[3] https://en.wikipedia.org/wiki/Personal_name
[4] https://www.wikidata.org

code for 193 countries in different languages, including English ($148,557$ unique family names). Since last names can cause gender and age ambiguity, especially if a name only consists of a single word which is a last name, the profiles that have only a family name ($100,741$ profiles) are removed from the data. In Table 1, the name 'Kim' contains a family name 'Kim' corresponding to the country Korea for both of them, so the profile is removed from the data to avoid gender ambiguity. The profile of 'Shady Decker' takes 'Shady' as its name excluding the last name 'Decker'.

After the cleaning procedure, $3,701,893$ profiles ($1,031,667$ unique names) remain. We create the dataset by using these remaining profiles. The final dataset consists of 1,745,437 items with name and demographic information.

3.3. Data Validation

To validate the dataset, we randomly selected 1,000 names from the prepared dataset. A research assistant and one of the authors manually checked the names for whether or not the name is likely to be a real name of a person. Both classified 996 names (99.6%) as real names. The four names not likely to be a person seem not the given names but nicknames 'Vanilla', 'Daddy', 'Strawberry' and 'Naruto' which are from Germany and United States where many profiles exist in the prepared dataset. Overall, this manual assessment suggests that the dataset is valid for the purposes of this study.

4. Algorithm Development

4.1. Gender and Country Authenticity

We begin by computing four basic indicators to measure if the name is gender, age, and country appropriate for a particular demographic group:

- **Demographic Frequency (DF):** The number of times the name appears among all profiles having the demographic. (Country Code, Age bin, Gender)
- **Country Frequency (CF):** The number of times the name appears among all profiles having the country code
- **Gender Frequency (GF):** The number of times the name appears among all profiles having the gender
- **Name Frequency (NF):** The number of times the name appears among all profiles

Even though one could possibly intuitively classify a name as being appropriate for a gender for a particular age and country, we cannot verify all the names by intuition for all age groupings and countries. To maximize the use of all the possibly usable names and to avoid choosing gender inappropriate names or family names, we apply a gender authenticity (GA) rating that is calculated for each name i based on the gender j as following:

$$GA_{ij} = \frac{GF_{ij}}{NF_i}$$

Where GF_{ij} is the frequency of the name i having the given gender j in the overall set of name entities. NF_i is the frequency of the name i in the overall set of name entities. A higher GA value of name for a specific gender indicates the name is popular in the specific gender and can be authentic for the gender. In turn, a mediocre GA value indicates the name can be gender ambiguous or can be a last name or a popular nickname used by both genders.

We also include a country authenticity (CA) rating. In many cases, the same name appears in different countries with various frequencies. However, certain names have significantly higher CF values for a particular country than the CF values for other countries. The names are more likely to be the authentic names in the specific country. To detect those names across countries, we quantify whether or not a specific name appears noticeably in specific countries by computing the Z-score. For the name i and the country j, the Z-score can be computed as follows:

$$Z_i^{(j)} = \frac{CF_i^{(j)} - avg(CF_i)}{\sigma}$$

Where CF_i is a set of frequencies having the given name for all the countries i. σ is the standard deviation of the CF_i. A higher Z-score indicates that the name is more likely to be used in the given country than in other countries. Then, using the Z-score value, we calculate the CA for the name i and country j as below:

$$CA_{ij} = \frac{CF_{ij}}{NF_i} \times Z_i^{(j)}$$

Where CF_{ij} is the frequency of the given name i for the given country j. NF_i is the frequency of the given name i in the entire profiles. A higher CA for a particular country indicates the name highly appears in the country relative to all countries. We use 1.0 when the Z-score is not available, since the given name i appears in a few countries.

4.2. Demonstration of Name Selection

The GA, CA, and F value of each name are useful indicators in selecting proper names for the given demographic information. By using GA, the algorithm can filter out the ambiguous names that are unlikely to be a given name. For instance, see Table 2, where 'Dream' can be simply filtered out by using the threshold of GA. When the DDP from Korea needs a name, 'Alex' can be excluded judging by its CA and F value even its GA value is above the threshold defined in Step 2. Accordingly, 'Alex' is most likely to be a name for the DDP from United States or Russia. 'Javad' and 'Lisa' are highly selectable names for their demographic groups by using their indicators.

4.3. Assigning Names to Entities

To assign a demographically appropriate name to entities such as DDPs, we develop an algorithm that applies the following steps:

Step 1: Find the corresponding names. This step takes country code, age bin, and gender as arguments. Then, it loads the dataset and finds the corresponding names that have the given demographic arguments.

Step 2: Select the names having the GA values greater than or equal to the threshold value 0.82, which is the mean GA value (SD = 0.27) of the names appearing at least twice in the profiles data ($NF > 1$).

Step 3: Calculate F for all names from Step 2. Then, choose the name with the highest F.

Step 4: Return the name as an appropriate name for the demographic information. Then assign the returned name to the corresponding DDP. Table 3 shows the name and demographic information for five personas after assigned names, as examples.

Some names have a high DF even though their CF and Z-score show that the names are not country specific. To avoid improper name selection for the country and demographics, we propose a weighted frequency F of the name i for the demographic j. F is calculated via the following equation:

$$F_{ij} = DF_{ij} \times CA_{ik}$$

Where DF_{ij} is the frequency of the name i for the given demographic information j. CA_{ik} is the authenticity of the name i for the given country information k.

5. Evaluation and Analysis

For evaluation, one needs to realize the approach we suggest is not based on prediction, but it is based on probabilistic calculations. Therefore, we do not apply evaluation metrics, such as accuracy or F1 score, that are typically used for evaluating machine learning models and algorithms. Instead, we evaluate the algorithm by developing two metrics:

- **Coverage Rate** – this is the proportion of found names from the evaluation dataset given the development set.
- **Demographic Appropriateness Score** – this is the proportion of names that native raters consider to be age, gender, and country appropriate.

5.1. Coverage Rate

InterPals was used for computing the F scores based on names and their associated demographics. Then, Wikidata is used for evaluating them. The inference chooses the demographic attributes with the highest F value for a particular name. The evaluation only uses the given names. We dropped some of the data tuples from the evaluation dataset due to duplication, being out of gender (not female or male), or age range being below the minimum (13–17). Among the remaining 841,249 instances, the coverage rate was 93.0%. Breakdown results are shown in Table 4.

5.2. Demographic Appropriateness

We evaluate the results with the help of native raters from different countries that were recruited among the authors' personal connections. To construct the evaluation dataset, we used stratified sampling to ensure a mix of names from different age groups and genders were represented from each country.

There were 33 raters for the 34 countries, obtained via convenience sampling from the authors' personal social networks. Each rater rated the names corresponding to their native country. One of the raters was a native of two countries, so he rated two countries. Eleven raters were women (33.3%).

Each rater was asked if the name represents a person in their country with the corresponding age group and gender: *"Does this name represent a person in your country with the indicated age range and gender?"*. The raters answered either 'yes' or 'no', and could leave comments if they wanted.

We obtained 1402 ratings from raters in 34 countries, representing an average of 41.2 ratings per country. We compute a demographic appropriateness score S by dividing the number of 'yes' answers (Y) of country i by the total number of answers (N) for the country i: $S = (Y_i/N_i) \times 100$

country_code	age_bin	gender	name	DF	GF	CF	NF	GA	Z-score	CA	F
IR	25-34	male	Javad	46	88	104	101	0.97	3.46	2.93	134.78
TH	18-24	female	Dream	33	87	127	182	0.7	6.46	3.09	101.91
	18-24	male	Dream	13	87	55	182	0.3	6.46	3.09	40.15
KR	45-54	male	Alex	8	84	11,868	13,856	0.86	0.0	0.0	0.0
RU	18-24	female	Lisa	243	451	4071	4123	0.99	4.03	0.44	107.07

Table 2. The example of names with their indicators

Name	Gender	Age	Country
Florencia	Female	34	Argentina
Tyler	Male	24	United States
Rahul	Male	32	India
Marie	Female	40	France
Aldrin	Male	30	Philippines

Table 3. Five DDPs with names assigned by the algorithm.

Demographic attribute	Coverage rate
Gender	89.74%
Country	41.77%
Age range	11.12%
All demographics	4.86%

Table 4. Breakdown of the demographic categories in the covered 93.0% of the evaluation data.

The results show an average S of 85.6%. Country-specific results are shown in Figure 2. Names from eight countries (23.5%) achieved a perfect score: Germany, Finland, United Kingdom, Italy, Japan, Lithuania, Poland, and United States.

The lowest score (10.5%) was obtained by Qatar, which is clearly an outlier compared to the rest of the countries. Switzerland was the second worst with 50.0%, but the score is nearly five times the score of Qatar. The low performance of Qatar can be explained due to cultural and population reasons – the native Qataris may not frequently visit international social networking sites like InterPals, and the population of Qatar consists of a high proportion of migrant workers.

Overall, the results indicate satisfactory performance, in that (a) a high number of names can be found in an independent dataset, and (b) most of the names the algorithm selects are gender, age and country appropriate.

5.3. Sub-group Analysis

We investigate the differences in the count of yes/no ratings across gender and age. A chi-square test of independence shows that there is no difference in the count of yes/no ratings between male and female groups, $X^2(1, N = 1402) = 2.16, p = .163$.

However, variation can be found by age, indicating that younger age groups tend to get better scores (see Figure 3). When using this insight to divide the data into two groups – Young (13-17, 18-24, 25-34) and Mature (45-54, 55-64, 65-) – we observe that the Mature group has significantly fewer 'yes' ratings, X^2 (1, N = 1201) = 75.27, $p < .00001$. This indicates that names for elderly demographics are more difficult to obtain.

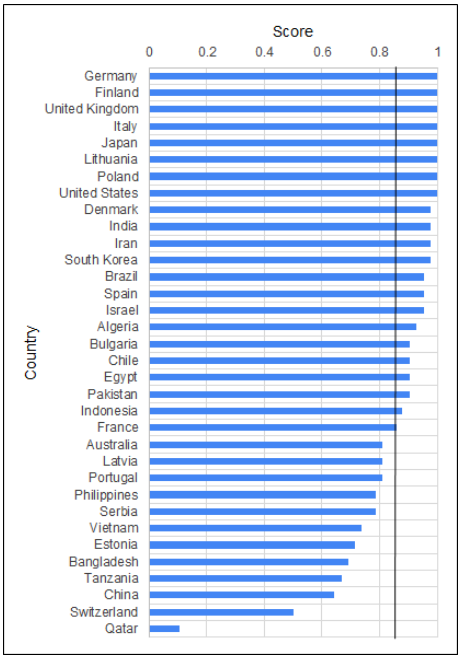

Figure 2. Evaluation scores for each country. The vertical line indicates average of all countries.

5.4. Failure Analysis

In some cases, the raters indicated a reason for not considering the name appropriate. One of the authors read the reasons and categorized them into seven groups. The groups, from most to least prevalent, include: *ethnically incorrect* (n = 24 instances), *wrong age* (n = 13), noise (n = 13), *wrong spelling* (n = 6), *full name* (n = 6), *nickname* (n = 4), and *wrong gender* (n = 1). Judging from these reasons, the demographic category that is the hardest to find names for is nationality, followed by age, then gender. Concerning the wrong spelling, some of the raters also considered the use of correct alphabets – for example, one rater pointed out that Jurg should be written as Jürg. Noise includes examples such as

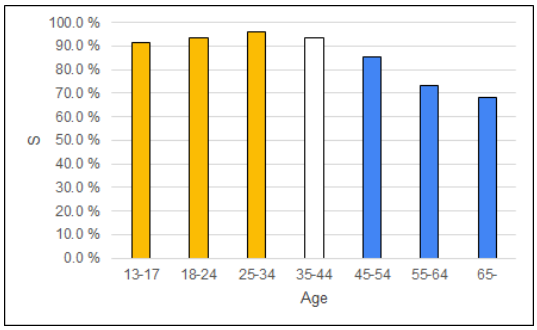

Figure 3. Demographic appropriateness scores (S) by age groups. Bars in orange color indicate the Young group and the blue color indicates the Mature group used in the chi-square test.

'Nordiclady', 'Ochendobraja', and 'Gkfkgfkgfkhfkh'.

6. Discussion

6.1. Main Contribution

The main contribution of this research is to propose a new method that automatically assigns demographically appropriate names to data-driven user segments with minimal manual human effort. We demonstrate the applicability of the method with the use case of DDPs. The method can also be implemented for personalized chatbots, virtual agents, or similar entities that require demographically appropriate names. Thus, the results extend beyond the context of DDPs – for example, to the humanization of chatbots [33].

The main advantage of the method relative to manual naming of the entities is that the statistical support of the demographic appropriateness is built into the decision-making of the algorithm. The higher the repetition of a particular name with a given demographic attribute, the higher the score assigned by the algorithm.

For example, if there are many "Jim's" in the dataset with age group of 45-54, gender of male, and country of United States, the algorithm will give a high score for "Jim" when selecting a name for this demographic combination. In contrast, manual selection of names – especially from countries that the selecting person is not a native – is constrained by the cultural knowledge of a person. While a person can pick reflective names for their home country, the selection is increasingly harder the more countries there are. Our method is beneficial in such cases.

6.2. Practical Implications

We release an online tool – **GAN2Name** – that can be used for name selection based on demographic attributes (age, gender, country). The tool can be accessed at [`https://quecst.qcri.org/tool/GAN2Name`]. In total, the tool covers $3,088$ demographic combinations. Most of the demographic combinations have a large variety of candidate names (M = 487.09, SD = $1,809$), which makes the tool useful for a system with broad representation of user segments. Increasing the F value to maximum decreases the number of available candidate names, but there still remain many candidates (M = 72.05, SD = 530.48) for system developers and designers to choose from.

6.3. Statement of Algorithmic Bias

Our approach favors majority names in the sense that the more a given name appears with a given demographics, the more likely it will be chosen. This means minority names might not be well presented, as some of the raters pointed out.

The German rater brought up caveats in his ratings: "*All the names listed are very German-national names (= high precision) but many living-in-Germany names are missing (= low recall). I couldn't find any Islamic/Turkish/Syrian name on the list, even though Mohammad is one of the most frequent male names in Germany. So the [names are] definitely biased and [represent] 'only found in Germany' names over 'typically found in Germany (and elsewhere)' names.*"

The rater from the United States also made a similar comment: *"(even though all the names were demographics appropriate,) the ethnic names missing. No Asians. No Mexicans. No 'stereotypical' Black names. All straight up Middle America White names."*

Also, there may be an effect of digital divide [34] in ethic minorities globally having less access to social media platforms and the Internet as a whole, as well as cultural reasons for not participating in online social networks [30]. Thus, bias can originate from two sources: the way the algorithm favors the majority (due its statistical nature) and due to potential under-representation of ethnic minority groups in the social media platform the data was collected from.

Finally, the binary notion of the gender is a worthwhile issue to point out. Most social media platforms still adhere to the binary concept of gender, not recognizing multiple gender identities. This means that any datasets collected from those platforms are also subject to this binary conceptualization of gender.

7. Conclusion, Limitations, and Future Work

We propose a methodology to assign demographically appropriate names to DDPs or

related entities such as virtual agents or chatbots. For this, we collected $3,817,272$ profiles from a large online community and created a dataset of $1,031,667$ unique names covering $3,088$ demographic combinations. We then defined *GA*, *CA*, and *F values* as indicators of authenticity for a demographic group. The proposed implementation returns demographically appropriate names while mitigating the need for manual effort.

There are several possibilities for improvement and future work. First, online users can have different nationalities from their resident countries, but all profiles do not have both country information. For example, in the Middle East, there can be many different ethnic groups across the region, including not only Arabs but Bengalis, Indians, Filipinos, and so on[5]. This can result in incorrect names for certain host countries. Among the prepared profiles, $592,746$ (16.01%) profiles have a hometown country information covering 70.14% of demographic combinations that the collected profiles can cover. If we could collect more reliable country information for names, the names would be representative of a given country.

Second, some countries do not have enough profiles given the data collection site, so the demographic groups of the countries are not sufficiently secured. On the other hand, the majority of the missing countries are listed as the smallest populations [6]. It is difficult to cover all the demographic groups for small population countries, as the population differences are likely reflected in the user base of the online social network.

Third, even if we remove as many family names from the collected names as possible, Wikidata is not a comprehensive dataset in this regard. Thus, there might still be family names that contain a group of words. To avoid losing the given names, Chinese and Korean names are skipped in the family name removal step when the name is a group of words and there is no connected words by hyphen. If we could find a way to classify the family name in the syllable based naming structure, we could obtain given names for the countries with a particular naming culture. Using only validated given names for the name selection could ensure more demographically appropriate name selection.

Fourth, the manual validation of the data results in the possibility that some popular nicknames in a certain country were considered as valid names. We also found some names not likely to be a given name but in some countries they are indeed given names, like 'Patience' and 'lord' in Ghana. The nuanced naming conventions in different countries require broad knowledge of culture to avoid stereotypical thinking.

Fifth, we used binary choice to evaluate if a name was appropriate. However, there could be borderline cases, in which the answer is somewhere in between. An alternative scale for evaluation would consider this and present the raters with multiple choices (e.g., "How appropriate, on a scale of 1–5, is this name?"). The evaluation could also include multiple raters per country to establish interrater reliability.

Finally, the source of names is an online social network where users can easily manipulate their names and use a nickname or pseudonym, not corresponding to their authentic name. To avoid such cases, manual filtering may be beneficial. From the results of the native evaluation, it appears that this "noise" contributes to the fact that the methodology did not achieve a perfect score for most of the countries, as there are some erratic names (see Section 5.4.). Future research could also address the inference of demographic attributes from noisy usernames.

[5] http://worldpopulationreview.com/continents/the-middle-east-population/

[6] https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

References

[1] I. Weber and A. Jaimes, "Demographic information flows," in *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 1521–1524, 2010.

[2] I. Weber and C. Castillo, "The demographics of web search," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, (New York, NY, USA), pp. 523–530, Association for Computing Machinery, July 2010.

[3] Y. Mejova, I. Weber, and L. Fernandez-Luque, "Online health monitoring using Facebook advertisement audience estimates in the United States: evaluation study," *JMIR public health and surveillance*, vol. 4, no. 1, p. e30, 2018. Publisher: JMIR Publications Inc., Toronto, Canada.

[4] M. Thelwall and E. Stuart, "She's reddit: A source of statistically significant gendered interest information?," *Information Processing & Management*, vol. 56, no. 4, pp. 1543 – 1558, 2019.

[5] J. An, H. Kwak, J. Salminen, S.-g. Jung, and B. J. Jansen, "Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data," *ACM Transactions on the Web (TWEB)*, vol. 12, no. 3, 2018.

[6] W. Huang, I. Weber, and S. Vieweg, "Inferring nationalities of Twitter users and studying inter-national linking," in *Proceedings of the ACM Conference on Hypertext and Social Media (HT'14)*, (Santiago, Chile), pp. 237–242, 2014.

[7] B. Bengtsson, J. K. Burgoon, C. Cederberg, J. Bonito, and M. Lundeberg, "The impact of anthropomorphic interfaces on influence understanding, and credibility," in *Proceedings of the Hawaii International Conference on Systems Sciences (HICSS)*, pp. 15–pp, 1999.

[8] A. Jenkinson, "Beyond segmentation," *Journal of targeting, measurement and analysis for marketing*, vol. 3, no. 1, pp. 60–72, 1994.

[9] B. J. Jansen, J. O. Salminen, and S.-g. Jung, "Data-Driven Personas for Enhanced User Understanding: Combining Empathy with Rationality for Better Insights to Analytics," *Data and Information Management*, vol. 4, pp. 1–17, Mar. 2020. Publisher: Sciendo Section: Data and Information Management.

[10] A. Cooper *et al.*, *The inmates are running the asylum:[Why high-tech products drive us crazy and how to restore the sanity]*, vol. 2. Sams Indianapolis, IN, 2004.

[11] H. Kwak, J. An, and B. J. Jansen, "Automatic Generation of Personas Using YouTube Social Media Data," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS-50)*, (Waikoloa, Hawaii), pp. 833–842, Jan. 2017.

[12] L. Nielsen, "From user to character: An investigation into user-descriptions in scenarios," in *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, pp. 99–104, 2002.

[13] S.-G. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona generation from aggregated social media data," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, (New York, NY, USA), pp. 1748–1755, ACM, 2017.

[14] J. An, H. Kwak, S.-g. Jung, J. Salminen, and B. J. Jansen, "Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data," *Social Network Analysis and Mining (SNAM)*, vol. 8, no. 1, 2018.

[15] J. Salminen, K. Guan, S.-G. Jung, S. A. Chowdhury, and B. J. Jansen, "A Literature Review of Quantitative Persona Creation," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'20)*, (Honolulu, Hawaii, USA), ACM, Apr. 2020.

[16] T. Mijač, M. Jadrić, and M. Ćukušić, "The potential and issues in data-driven development of web personas," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1237–1242, May 2018.

[17] J. Salminen, L. Nielsen, S.-G. Jung, J. An, H. Kwak, and B. J. Jansen, ""Is More Better?": Impact of Multiple Photos on Perception of Persona Profiles," in *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI2018)*, (Montreal, Canada), Apr. 2018.

[18] L. Nielsen, K. S. Hansen, J. Stage, and J. Billestrup, "A template for design personas: analysis of 47 persona descriptions from danish industries and organizations," *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, vol. 7, no. 1, pp. 45–61, 2015.

[19] E. Kušen and M. Strembeck, "You talkin' to me? exploring human/bot communication patterns during riot events," *Information Processing & Management*, vol. 57, no. 1, p. 102126, 2020.

[20] A.-K. Cordes, B. Barann, M. Rosemann, and J. Becker, "Semantic Shopping: A Literature Study," in *Proceedings of the Hawaii Conference on System Sciences (HICSS)*, 2020.

[21] J. H. Kim and Y. Kim, "Instagram user characteristics and the color of their photos: Colorfulness, color diversity, and color harmony," *Information Processing & Management*, vol. 56, no. 4, pp. 1494 – 1505, 2019.

[22] B. Hofstra and N. C. de Schipper, "Predicting ethnicity with first names in online social media networks," *Big Data & Society*, vol. 5, no. 1, 2018.

[23] K. Kalimeri, M. G. Beiró, M. Delfino, R. Raleigh, and C. Cattuto, "Predicting demographics, moral foundations, and human values from digital behaviours," *Computers in Human Behavior*, vol. 92, pp. 428–445, 2019.

[24] K. Li, L. Cheng, and C.-I. Teng, "Voluntary sharing and mandatory provision: Private information disclosure on social networking sites," *Information Processing & Management*, vol. 57, no. 1, p. 102128, 2020.

[25] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users," in *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, 2011.

[26] W. Liu and D. Ruths, "What's in a Name? Using First Names as Features for Gender Inference in Twitter," in *Analyzing Microtext AAAI 2013 Spring Symposium*, vol. 13, pp. 10–16, 2013.

[27] E. Cunha, G. Magno, M. A. Gonçalves, C. Cambraia, and V. Almeida, "He Votes or She Votes? Female and Male Discursive Strategies in Twitter Political Hashtags," *PLOS ONE*, vol. 9, p. e87041, Jan. 2014.

[28] J. Mueller and G. Stumme, "Gender Inference using Statistical Name Characteristics in Twitter," in *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on Social Informatics*, MISNC, SI, DS 2016, (Union, NJ, USA), pp. 1–8, Association for Computing Machinery, Aug. 2016.

[29] H. Oktay, F. Aykut, and Z. Ertem, "Demographic Breakdown of Twitter Users: An analysis based on names," in *Proceedings of the Academy of Science and Engineering (ASE)*, June 2014.

[30] J. Salminen, S. Şengün, H. Kwak, B. J. Jansen, J. An, S.-g. Jung, S. Vieweg, and F. Harrell, "From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas," *First Monday*, vol. 23, June 2018.

[31] A. T. Nguyen, M. Lease, and B. C. Wallace, "Explainable modeling of annotations in crowdsourcing," in *Proceedings of the ACM Conference on Intelligent User Interfaces (IUI'19)*, pp. 575–579, 2019.

[32] S. Lomborg and A. Bechmann, "Using APIs for data collection on social media," *The Information Society*, vol. 30, no. 4, pp. 256–265, 2014. Publisher: Taylor & Francis.

[33] E. Go and S. Shyam Sundar, "Humanizing Chatbots: The effects of visual, identity and conversational cues on humanness perceptions," *Computers in Human Behavior*, Jan. 2019.

[34] J. Pick and A. Sarkar, "Theories of the digital divide: Critical comparison," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, pp. 3888–3897, IEEE, 2016.