# Beyond names: how to label gender automatically in CMC data?

**Pasi Fränti, Juhani Järviö, Mehrdad Salimi, Irene Taipale, Mikko Laitinen, Rahel Albicker, Chunyan Nie, Masoud Fatemi, Paula Rautionaho**

University of Eastern Finland

E-mail: mikko.laitinen@uef.fi

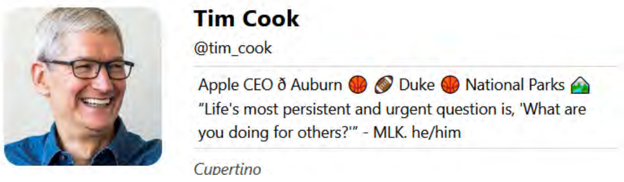
### Abstract

Large-scale data from social media offers numerous benefits for research, but one significant and widely known limitation is the lack of detailed social background information. This gap poses a serious challenge for fields such as sociolinguistics and the study of language variation and change, where demographic and contextual information are crucial. A commonly used approach in computer-mediated communication (CMC) research has been to infer gender from users' names. However, a variety of other methods have emerged in recent years, drawing from advances in machine learning. This presentation reviews the current state of social media data enrichment and introduces a generalizable method that integrates various types of background information. Enriched data can train machine learning models to label social media user accounts with more accurate gender information.

**Keywords:** demographic prediction, Twitter data, data labeling, social media

## 1. Introduction

Social media datasets are widely used in sociolinguistic research, yet they typically lack the most basic demographic variables, such as age, gender, or occupation.[1] As a result, socially conditioned variation remains largely inaccessible in studies that draw on large-scale social media corpora. We use datasets of 343,149 Twitter users from Australia, the UK, and the USA, with message histories from 2006 to 2023 (Laitinen et al., 2025).[2] So far, we have enriched the data with network information (Laitinen & Fatemi, 2024), and our current focus is on inferring gender information from user profiles. While this application does not provide a standardized format for users to express their gender, many profiles contain sufficient direct or indirect cues to make gender inference feasible. To illustrate, Figure 1 provides information pointing toward a male gender identity. These include a profile picture, a traditionally male first name in the Anglo-American culture, and a pronoun declaration (*he/him*).



**Tim Cook**
@tim_cook

Apple CEO ð Auburn 🏀 🏈 Duke 🏀 National Parks 🏞
"Life's most persistent and urgent question is, 'What are you doing for others?'" - MLK. he/him

*Cupertino*

Figure 1: Example profile of a known public figure

This is not always the case with our dataset, which is also too large to annotate manually. If it took a human 10 seconds to infer the gender of a user, the full annotation task would require around 1,000 hours of work for a single person. Many user profiles also lack clear indicators, such as pictures or names, and may contain contradictory information, making them much more difficult to annotate. For these reasons, we rely on computational methods to assist with labeling.

Previous studies have employed a variety of methods, ranging from using users' names (Coats, 2016, 2019) and profile pictures (Yildiz et al., 2017) to analyzing textual information in user profiles (Wang et al., 2019). A commonly used method in computer-mediated communication (CMC) studies is the use of names, as first-name databases are widely available across many languages (Lupo et al., 2024). For example, the UK Office for National Statistics (ONS) provides data on names given to newborns between 1996 and 2021, along with associated gender information. These databases are generally consistent, with only a small minority of names assigned to more than one gender. However, naming practices are highly heterogeneous, and a significant share of accounts may remain unlabeled. This article surveys the state of the art to support more accurate labeling.

We mostly focus on binary classification of gender for the results. This is a practical decision, as most methods that have high coverage only make binary predictions: either male or female. Three of the methods we consider also offer a non-binary category: pronoun declarations, keyword search, and crowdsourcing. However, none of them has enough coverage to be useful on a large scale. We acknowledge the limitations of binary gender classification, which does not capture the complexity and diversity of identities (Keyes et al., 2021; Simeoni et al., 2024). However, due to the scale of the dataset and the constraints of current computational methods, we settled on binary categorization at this stage.

[1] We wish to thank the two anonymous reviewers for their comments on an earlier version of this article.

[2] Twitter's name was changed to X recently. However, our datasets were collected before this change. For consistency and accuracy, we use the original name to refer to our dataset.

## 2. Previous studies

Gender (Yildiz et al., 2017) and age (Sloan et al., 2015) prediction have received considerable attention in recent research, with Twitter serving as the primary data source (Lupo et al., 2024; Mislove et al., 2011; Sloan et al., 2015; Tonglet et al., 2024; Yildiz et al., 2017). Previous studies have also focused on predicting other demographic factors, such as location (Lupo et al., 2024; Sloan et al., 2013; Wang et al., 2019), occupation (Sloan et al., 2015), and social class (Sloan et al., 2015). However, some demographic attributes are more readily predictable than others. For example, it has been argued that age is easier to predict than occupation (Sloan et al., 2015) but more challenging than gender (Lupo et al., 2024; Wang et al., 2019).

A key challenge in using social media data is its lack of demographic representativeness, particularly concerning geography, gender, and race (Mislove et al., 2011; Wang et al., 2019). There is a known bias toward males in Twitter data compared to actual populations (Dixon, 2024; Gombert et al., 2025; Mislove et al., 2011; Yildiz et al., 2017). Twitter users are also, on average, significantly younger than the population at large (Sloan et al., 2015), which affects the ability to generalize social media results. Demographic prediction offers a potential means to mitigate this limitation. When demographic information is available, researchers can address sampling biases such as overrepresentations of young men.

In manual labeling, annotators typically rely on a combination of profile-based signals such as the username, display name, profile description, and profile picture. Message content has also been used to infer user demographics (Yildiz et al., 2017). The scale of annotation varies across studies. For instance, in Wang et al. (2019), the labeling was conducted by three individuals, while Yildiz et al. (2017) involved over 1,000 annotators. Manual labeling allows nuanced interpretation of multiple cues with high accuracy but is time-consuming and does not scale well.

Automatic labeling methods have been developed to address these challenges. Common approaches include image processing based on profile pictures or searching for gendered lists of names and keywords. Some studies have also applied regular expressions to users' profile descriptions to detect age (Lupo et al., 2024; Sloan et al., 2015). These methods are scalable. However, some of these have limited coverage; e.g., a face could be detected only in about 30% of the profile pictures (O'Connor et al., 2024).

The effectiveness of these approaches varies depending on the task and data quality. Crowdsourcing has been reported to work better for age detection than computer-based methods (Yildiz et al., 2017), whereas another more recent study reports that computer-based methods can already predict gender with high accuracy (Wang et al., 2019).

The most common approach to gender prediction is to use lists of first names (Sloan et al., 2013). In this approach, names are classified based on the probability of being associated with one gender. For example, if 98 of 100 babies named *Lennox* are male and two are female, the probability of the name being male is 98%. Following Mislove et al. (2011), we use a 95% probability threshold for choosing names. However, accuracy varies between languages. For instance, Italian names are more gender-specific, allowing for more reliable prediction (Lupo et al., 2024). In general, this approach does not provide high coverage: Mislove et al. (2011) managed to predict the gender of only 64.2% of users.

Machine learning models based on text have also been used for gender prediction. Lupo et al. (2024) consider Bag of Words, TF-IDF, word embeddings, and topic modeling for feature extraction. Using the M3 inference tool, Gonzales (2024) utilizes grammatical features for age and gender prediction. Large language models (LLMs) and machine learning classifiers such as Random Forest and XGBoost have been applied with promising results (Tonglet et al., 2024; Wang et al., 2019). Tonglet et al. (2024) report 92% accuracy for gender detection with M3 for 82% of users in their dataset. While these methods can leverage different data like text and images, they have comparability issues due to a lack of standardized evaluation datasets (O'Connor et al., 2024).

A challenge in gender and age prediction is the lack of standardized reporting on data collection and labeling procedures (O'Connor et al., 2024). We aim to address this by providing a detailed account of our data collection and annotation process. Our computer software is publicly available as open source to support transparency and replication (see section 6 below).

## 3. Data

Our data consists of profile information and user-generated messages of 343,149 Twitter users, mostly from the USA, UK, and Australia. Data gathering is explained in more detail in Laitinen et al. (2025). The profiles were collected in 2023 using the now-closed Academic API, which limited the data collection to a maximum of 3,200 messages per user, spanning from 2006 to 2023. Network data was also collected based on ego networks, centered around 5,773 ego users. However, we focus here on individual profiles for predicting gender and aim at combining network data with gender/age data later.

The profile information includes username, display name, description, and location fields. All of these were used for the methods described in the following sections. The location field is not useful as an actual geographical datapoint, since it is a free-text field, and according to Mislove et al. (2011), only 9% of self-announced user locations can be matched to a real location by the Google Maps API. Because of the age of the data and the changes in the API, retrieving profile images at a massive scale is difficult, meaning that images were not used in this paper.

User language is available for each post, provided by the Twitter API, and is used to estimate each user's primary language. In our data, primary language means that at least

75% of the user's tweets were in a single language. Table 1 shows that 82% of users write mainly in English.

| Language | Users | |
|---|---|---|
| English | 280,044 | 82 % |
| Multiple | 46,915 | 14 % |
| Spanish | 3,493 | 1 % |
| Japanese | 2,190 | 1 % |
| No linguistic content | 1,470 | 0 % |
| **Top 5 languages** | **334,112** | **97 %** |
| **All languages** | **343,146** | **100 %** |

Table 1: Top 5 languages as the number of users whose tweets were 75% in the listed language.

Information on user location is limited as it is available only for a subset of tweets. Only about one-third of users have any country-level location data. Of those, 83% are from the three countries from which the original data was gathered: the US, UK, and Australia.

To fix the gaps in the profile data, we set up a crowdsourcing effort to manually label a small subset. This crowdsourcing used a combination of profile data and user-generated messages as the stimuli for our students, who manually labeled some 2,800 accounts. This data will be used as the gold standard to evaluate the accuracy of automated labelling.

## 4. Parameters for computer-based labelling

For automatic labeling, we use a combination of three parameters:

- Name information
- Self-declaration of gender
- Keywords in user profiles

These methods were chosen for their simplicity and presumably high accuracy for automated labeling.

### 4.1. Name-based classification

One of the most widely used methods is to tokenize usernames and match the output against a priori collected first name lists with gender information (Coats, 2016; Mislove et al., 2011; Sloan et al., 2013).

We acquired first name lists from four national government organizations: the UK, US, Australia, and Canada to increase accuracy (Attorney-General's Department, Government of South Australia, 2013; Office for National Statistics, 2022; Statistics Canada, 2023; U.S. Social Security Administration, 2024). Their statistics are summarized in Table 2.

| Region | Years | Names | Average M\|F confidence | |
|---|---|---|---|---|
| UK | 1996–2021 | 36,043 | 99 % | 99 % |
| USA | 1923–2023 | 101,785 | 98 % | 99 % |
| Australia | 1944–2014 | 51,518 | 99 % | 99 % |
| Canada | 1991–2023 | 16,444 | 98 % | 99 % |
| **Unique names** | **M:43,441** | **119,244<br>F:75,655** | **98 %<br>M** | **99 %<br>F** |

Table 2: Summary of the name datasets. Average confidence is the probability that a name is a specific gender, averaged across all names.

We also employ a list of last names from the US 2010 census data (U.S. Census Bureau, 2016) to filter out names that are predominantly surnames, even if they are occasionally used as first names. One such example is *Williams*, which occurs 1,625,252 times as a surname compared to just 5,295 times as a first name.

The datasets contain names, counts, and the gender information (male or female) for each name, excluding names with <10 instances. Names that can be given both to girls and boys add uncertainty to the method (e.g., *Riley*, with 51% male). We tokenize usernames and profile names using the NLTK natural language toolkit Python package, along with various manual fixes to parse more difficult names. We match each token to the combined name list and, in total, found a match for 275,143 users (80%), of which 152,612 (45%) were gendered based on first names with an average probability of one gender >95%.

Mislove et al. (2011), using the 1,000 most common male and female names from the Social Security Administration dataset, found a match for 64% of users, of which 72% had a male name. Sloan et al. (2013) used a database of 40,000 names from 54 countries globally. They found a first name for 48% of the users. Both Mislove and Sloan used only the first name item.

### 4.2. Pronoun declarations

The second approach builds on an increasing tendency of users to self-declare their gender identity in the profile info (e.g., *she/her*). Recent studies show that this is an emerging way of identifying oneself (Jiang et al., 2023; Tucker & Jones, 2023).

We found this method to be both the most reliable and the easiest to detect. Self-declaration also extends to non-binary genders.

To implement it, we developed a list of regular expressions designed to match known English pronoun declarations. This list includes 200 variations, with the most frequent ones shown in Table 3. Remarkably, 94% of all pronoun declarations fall into the ten most common patterns. To identify these declarations, we tokenized the text by extracting words separated by standard delimiters and matched them against our list. All matches were manually reviewed to eliminate false positives. Each declaration was then categorized into one of four groups: female (F), male (M), non-binary (NB), or uncategorized (U).

| Declaration | | Count | |
|---|---|---|---|
| She / her | F | 6,178 | 45 % |
| He / him | M | 3,519 | 26 % |
| They / them | NB | 926 | 6 % |
| She / they | F | 835 | 6 % |
| He / they | M | 430 | 3 % |
| **Top 5** | | **11,888** | **91 %** |
| **All** | | **13,447** | **100 %** |

Table 3: Top 5 most common pronoun declarations. (Taken from a total 343k user dataset.)

The limitation is low coverage, as only about 5% of users in our data include pronoun declarations. The method is also limited to English-language pronouns. The main advantage of pronoun declaration is its high accuracy.

Jiang et al. (2023) reported an increase of 33% in the number of tweets made by users with pronouns from 2020 to 2021 (2.86% → 3.82%). Tucker & Jones (2023) write that pronoun usage peaked by 2022 to around 4%–5%. Our results show the same for profiles fetched in 2023. Both studies also show a similar distribution of gender categories to ours, with female pronoun declarations being the majority. Table 4 shows the gender distribution in pronoun declarations.

| Unique users | Count | |
|---|---|---|
| Female | 7,445 | 55 % |
| Male | 4,359 | 32 % |
| Non-binary | 962 | 7 % |
| Unknown | 681 | 5 % |
| **Total** | **13,447** | **100 %** |

Table 4: Counts of users with pronoun declarations. The coverage remains relatively low: 13,447 / 343,149 = 4%.

### 4.3. Keyword search

The third approach is to find gender-related keywords in the profile description (Emmery et al., 2017; Jurgens et al., 2017). We experimentally compiled three keyword lists targeting gendered terms associated with female, male, and non-binary identities (e.g., *mother*, *husband*, and *enby*). These keywords were used in regular expression-based searches of users' profile descriptions to identify gender-related self-descriptions. Unfortunately, like pronoun declarations, keyword search also has low coverage, with only 6% of profiles labelled as one of the three gender categories: F, M, NB.

## 5. Observations

The gender labelling task has two objectives:

1. *Coverage*: to label as many profiles as possible
2. *Accuracy*: to ensure the labels are accurate

We measure the success of the different methods by their coverage and accuracy. *Coverage* is the number of users that are assigned a label, while *accuracy* is the proportion of labeled users for whom the predicted gender matches the true gender. Both are reported as a proportion between 0 and 100%.

In computer science, their joint optimization is known as a two-objective optimization problem. Perfect results for both metrics are rarely achievable simultaneously: improving one typically compromises the other.

Reducing the problem to a single-objective optimization, a simple goal would be to label all users (100% coverage) and find a method to maximize accuracy. It would also be possible to achieve high accuracy by labelling only the users that we are certain of. This could give close to 100% accuracy at the cost of low coverage.

In practice, perfect accuracy is unattainable even for a subset of users. Automatic methods can be surprisingly good but are far from perfect. Even humans may fail, as user profiles may lack obvious gender-related clues such as a name and profile picture. A realistic goal is therefore to find a good compromise by giving more weight to the more important objective.

Next, we report the observed values for the two objectives when labeling a subset of our dataset using three automatic methods. We lack ground truth labels and therefore use the human-annotated genders as a gold standard.

The results are reported in Table 5. Our first observation is that names offer the widest coverage (44%) but have the lowest accuracy (72%). In contrast, pronoun declaration and keyword-based methods show higher accuracy (96% and 82%) but considerably lower coverage (3% and 6%).

Table 5 also includes results obtained with the M3 method (Wang et al., 2019), excluding predictions based on profile pictures. M3 uses sophisticated machine learning techniques with word embedding. It achieves higher coverage (63%), but lower accuracy (65%) compared to name-based predictions alone (72%). Its accuracy can be increased from 65% to 79% by raising its confidence threshold from 50% to 95%, although this reduces coverage to 28%. Including profile pictures would increase accuracy, but as explained above, we did not use them for this test due to the scale of the data.

We also note that the methods display gender-related biases, as the proportion of users classified as female is systematically higher with the automatic methods than in the manually annotated sample. For example, name-based and keyword-based predictions identify 43% and 46% users as female, respectively. Women seem to use pronoun declarations more often than men in their profiles, as 63% of the detected declarations are female. This result differs considerably from the usual male-female ratio reported elsewhere. For example, Dixon (2024) reported 60% male, and our crowdsourcing tool shows 70% male.

| Feature | Female labels | Coverage | Estimated accuracy |
|---|---|---|---|
| Names | 43 % | 152,612 | 72 % |
| Pronouns | 63 % | 11,729 | 96 % |
| Keywords | 46 % | 22,248 | 84 % |
| Manual | 35 % | 1,888 | - |
| M3 (50 %) | 38 % | 217,563 | 65 % |
| M3 (95 %) | 26 % | 96,322 | 79 % |

Table 5: Results table from several methods combined. Manual labelling was used to measure estimated accuracy. M3 was split into two: one with a simple majority (the bigger probability is the label), and one with a 95% threshold, where any users with a lower probability were labeled as undetermined.

## 6. Conclusions and next steps

Incorporating social background information into large-scale CMC data may initially appear straightforward. However, several factors make this a complex and nuanced task. From a social sciences and humanities perspective, many social background variables, such as gender, age, and class, are not fixed traits but socially constructed categories. Individuals perform and express these identities in diverse ways. This complexity is especially evident on social media, where users curate and perform their identities through varied and often ambiguous signals.

From a computational perspective, the challenge becomes methodological: how can we ensure both high coverage and high accuracy in labeling? Prioritizing high accuracy reduces coverage, while maximizing coverage compromises precision. Effective research must strike a balance between these competing goals. Our approach addresses this by constructing a dataset that integrates multiple parameters to support robust demographic inference.

This article reviews the state-of-the-art in socio-demographic user labeling, drawing on recent advances not only in CMC research but also in machine learning and related fields. Building on interdisciplinary work, we outline a methodology for constructing a testing dataset that incorporates three gender-related parameters.

We also built a crowdsourcing platform, designed for flexible task configuration and adaptable for a wide range of annotation needs. This platform underpins the creation of our testing dataset, which will be our ground truth. It can be used to evaluate our automatic methods, and other ML-based methods, such as the M3 model (Wang et al., 2019).

Our immediate goal is to complete gender labeling, aiming to maximize coverage while maintaining acceptable accuracy for all 343,149 users in the dataset. In the future, we plan to extend our work to include additional demographic variables, such as age groups (e.g., <29 years, 30–50 years, 51+). Notably, our dataset already includes network parameters that capture the degree of user connectivity. When combined with gender and age data, this enables large-scale sociolinguistic analysis.

The resulting data will be made public for reuse and further research once we have annotated all our data. The current code for the presented automatic labeling is available at https://cs.uef.fi/comet/code/.

## 7. References

Attorney-General’s Department, Government of South Australia. (2013). *Most popular Baby Names (1944-2013)* (Version 2016) [Statistical dataset; CSV]. data.sa.gov.au. https://data.sa.gov.au/data/dataset/popular-baby-names

Coats, S. (2016). Grammatical frequencies and gender in Nordic Twitter Englishes. In *Proceedings of the 4th Conference on CMC and Social Media Corpora for the Humanities*. University of Ljubljana Academic Publishing, pp. 12–16 https://nl.ijs.si/janes/wp-content/uploads/2016/09/CMC-2016_Coats_Grammatical-Frequencies-and-Gender.pdf

Coats, S. (2019). Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics*, 42(1). https://doi.org/10.1017/S0332586519000039

Dixon, S. J. (2024, May 22). *Distribution of X (formerly Twitter) users worldwide as of January 2024, by gender*. Statista. https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/

Emmery, C., Chrupała, G., & Daelemans, W. (2017). Simple Queries as Distant Labels for Predicting Gender on Twitter. *Proceedings of the 3rd Workshop on Noisy User-Generated Text*, 50–55. https://doi.org/10.18653/v1/W17-4407

Gombert, A., Sánchez-López, B., & Cerquides, J. (2025). Jekyll institute or Mrs Hyde? Gender identification with machine learning. *Engineering Applications of Artificial Intelligence*, *144*, 110087. https://doi.org/10.1016/j.engappai.2025.110087

Gonzales, W. D. W. (2024). When to (not) split the infinitive: Factors governing patterns of syntactic variation in Twitter-style Philippine English. *English Language & Linguistics*, 28(2), 305–339. https://doi.org/10.1017/S1360674323000631

Jiang, J., Chen, E., Luceri, L., Murić, G., Pierri, F., Chang, H.-C. H., & Ferrara, E. (2023). *What are Your Pronouns? Examining Gender Pronoun Usage on Twitter*. https://doi.org/10.36190/2023.02

Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). Writer Profiling Without the Writer’s Text. In G. L. Ciampaglia, A. Mashhadi, & T. Yasseri (Eds.), *Social Informatics*

(Vol. 10540. Springer International Publishing, pp. 537–558 https://doi.org/10.1007/978-3-319-67256-4_43

Keyes, O., May, C., & Carrell, A. (2021). You Keep Using That Word: Ways of Thinking about Gender in Computing Research. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 39, pp. 1–23. https://doi.org/10.1145/3449113

Laitinen, M., & Fatemi, M. (2024). Testing the weak-tie hypothesis with social media. *11th Conference on Computer-Mediated Communication and Social Media Corpora*, pp. 46–51. https://shs.hal.science/halshs-04673776/file/241007_CMC_Proceedings_DOI.pdf#page=60

Laitinen, M., Rautionaho, P., Fatemi, M., & Halonen, M. (2025). Do we swear more with friends or with acquaintances? F#ck in social networks. *Lingua*, *320*, 103931. https://doi.org/10.1016/j.lingua.2025.103931

Lupo, L., Bose, P., Habibi, M., Hovy, D., & Schwarz, C. (2024). *DADIT: A Dataset for Demographic Classification of Italian Twitter Users and a Comparison of Prediction Methods* (No. arXiv:2403.05700). arXiv. https://doi.org/10.48550/arXiv.2403.05700

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2011). Understanding the Demographics of Twitter Users. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1), Article 1. https://doi.org/10.1609/icwsm.v5i1.14168

O'Connor, K., Golder, S., Weissenbacher, D., Klein, A. Z., Magge, A., & Gonzalez-Hernandez, G. (2024). Methods and Annotated Data Sets Used to Predict the Gender and Age of Twitter Users: Scoping Review. *Journal of Medical Internet Research*, 26(1), e47923. https://doi.org/10.2196/47923

Office for National Statistics. (2022). *Baby names in England and Wales: From 1996* (Version 2022) [Statistical dataset; XLSX]. Office for National Statistics (ONS). https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/livebirths/datasets/babynamesinenglandandwalesfrom1996

Simeoni, F., Menéndez-Blanco, M., Vyas, R., & De Angeli, A. (2024). Querying the Quantification of the Queer: Data-Driven Visualisations of the Gender Spectrum. *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pp. 3243–256. https://doi.org/10.1145/3643834.3660695

Sloan, L., Morgan, J., Burnap, P., & Williams, M. (2015). Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLOS ONE*, 10(3), e0115545. https://doi.org/10.1371/journal.pone.0115545

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, 18(3), 74–84. https://doi.org/10.5153/sro.3001

Statistics Canada. (2023). *First names at birth by sex at birth, selected indicators* [Dataset]. Government of Canada. https://doi.org/10.25318/1710014701-ENG

Tonglet, J., Jehoul, A., Reusens, M., Reusens, M., & Baesens, B. (2024). Predicting the demographics of Twitter users with programmatic weak supervision. *TOP*, *32*(3), pp. 354–390. https://doi.org/10.1007/s11750-024-00666-y

Tucker, L., & Jones, J. (2023). Pronoun Lists in Profile Bios Display Increased Prevalence, Systematic Co-Presence with Other Keywords and Network Tie Clustering among US Twitter Users 2015-2022. *Journal of Quantitative Description: Digital Media*, *3*. https://doi.org/10.51685/jqd.2023.003

U.S. Census Bureau. (2016). *Frequently Occurring Surnames from the 2010 Census* [Dataset]. U.S. Census Bureau; U.S. Census Bureau. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

U.S. Social Security Administration. (2024). *Baby Names from Social Security Card Applications—National Data* [CSV]. U.S. Social Security Administration. https://catalog.data.gov/dataset/baby-names-from-social-security-card-applications-national-data

Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019). Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *Proceedings of the 2019 World Wide Web Conference*, pp. 2056–2067. https://doi.org/10.1145/3308558.3313684

Yildiz, D., Munson, J., Vitali, A., Tinati, R., & Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, 37(46), pp. 1477–1514. https://doi.org/10.4054/DemRes.2017.37.46