Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM 2023)

Large-Scale Demographic Inference of Social Media Users in a Low-Resource Scenario

Karim Lasri[1, 2 *], **Manuel Tonneau**[1, 3 *], **Haaya Naushan**[1], **Niyati Malhotra**[1], **Ibrahim Farouq**[1, 4], **Victor Orozco-Olvera**[1], **Samuel Fraiberger**[1, 5, 6]

[1] The World Bank,
[2] Ecole Normale Supérieure - PSL,
[3] University of Oxford,
[4] Universiti Sultan Zainal Abidin,
[5] New York University,
[6] Massachusetts Institute of Technology
{klasri, mtonneau, hnaushan, nmalhotra, ifarouq, vorozco, sfraiberger}@worldbank.org

Abstract

Characterizing the demographics of social media users enables a diversity of applications, from improved targeting of policy interventions to the derivation of representative population estimates of social phenomena. Achieving high performance with supervised learning, however, can be challenging as labeled data is often scarce. Alternatively, rule-based matching strategies provide well-grounded information but only offer partial coverage over users. It is unclear, therefore, what features and models are best suited to maximize coverage over a large set of users while maintaining high performance. In this paper, we develop a cost-effective strategy for large-scale demographic inference by relying on minimal labeling efforts. We combine a name-matching strategy with graph-based methods to map the demographics of 1.8 million Nigerian Twitter users. Specifically, we compare a purely graph-based propagation model, namely Label Propagation (LP), with Graph Convolutional Networks (GCN), a graph model that also incorporates node features based on user content. We find that both models largely outperform supervised learning approaches based purely on user content that lack graph information. Notably, we find that LP achieves comparable performance to the state-of-the-art GCN while providing greater interpretability at a lower computing cost. Moreover, performance does not significantly improve with the addition of user-specific features, such as textual representations of user tweets and user geolocation. Leveraging our data collection effort, we describe the demographic composition of Nigerian Twitter and find that it is a highly non-uniform sample of the general Nigerian population.

1 Introduction

With half of the world's population regularly logging in, online social media has become central to our daily lives as a tool for communication and information. This global uptake has also made social media an important source of data to study social phenomena, from complementing imperfect official statistics through *social sensing* (Llorente et al. 2015; Kryvasheyeu et al. 2016; Palotti et al. 2020) to studying the offline social impacts of this new technology (Lorenz-Spreen et al. 2023). An important limitation of social media data for social science research is that for most platforms and in most countries, especially in the Global South, the demographic makeup of the social media population is unknown. Therefore, accurately predicting attributes of social media users is key to performing in-depth analyses of online social phenomena, such as understanding interactions between groups of social media users or examining the representativeness of a given user population (Tufekci 2014).

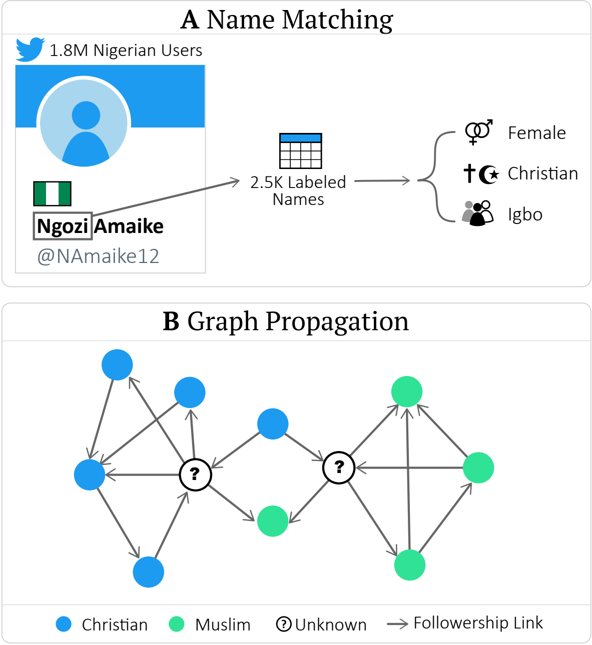


Figure 1: Illustration of the key method used in this paper

Over the past several years, a number of methods have been developed to accurately infer the demographic characteristics of social media users. Previous research in this field used supervised learning, relying on labeled user data

[*] Equal contribution.

and deriving features from user metadata, text and network information (Preoţiuc-Pietro, Lampos, and Aletras 2015; Chakraborty et al. 2017; Preoţiuc-Pietro and Ungar 2018; Wood-Doughty et al. 2018; Pan et al. 2019; Wang et al. 2019). When labeled user data is unavailable, external data sources, occasionally combined with rule-based methods, have successfully been employed to infer demographics (Mohammady and Culotta 2014; Culotta, Kumar, and Cutler 2015; Mislove et al. 2011; Liu and Ruths 2013; Karimi et al. 2016). These methods have been applied at scale to map targeted characteristics of social media populations (Mislove et al. 2011; Sloan et al. 2015; Mellon and Prosser 2017). This vast literature relies on a wide array of models and input features; yet, it is unclear which combination generalizes best for population-level descriptions. This is especially true for deployment over a large, unfiltered set of users with limited labeled resources. In this context, rule-based approaches, such as name matching, are often the only cost-effective option but lack coverage.

A promising solution to increase coverage under low-resource constraints is to draw insights from available network information using graph-based techniques. These approaches leverage social media connectivity found in followership or user-to-user interactions to propagate information from labeled to unlabeled users (Wang and Zhang 2006; Kipf and Welling 2017; Gao, Wang, and Ji 2018). While graph propagation proved successful for demographic inference (Speriosu et al. 2011; Li, Xu, and Lu 2015; Kim et al. 2017; Pan et al. 2019), it is yet to be determined what method is most effective with limited resources.

In this paper, we investigate the potential of graph-based propagation in combination with domain knowledge from name-based dictionaries to perform robust demographic inference at scale in a low-resource context. To this end, we evaluate two graph-based propagation models, namely Label Propagation (LP) and Graph Convolutional Neural Network (GCN), on Nigerian Twitter. We find that both LP and GCN largely outperform rule-based and supervised learning approaches while achieving full coverage, even on accounts with low followership or user content. Text-based supervised classification methods do not perform well in this context as the majority of users produce little content. Additionally, we find that LP achieves performance levels similar to the more recent GCN, while being more interpretable. We also find that incorporating user-profile features into a GCN does not improve performance significantly. Finally, we describe the demographic composition of the Nigerian Twitter population and find that it is a highly non-uniform sample of the general population. We therefore make the following contributions:

- A detailed comparison of scalable demographic inference methods in a low-resource context.
- A description of the demographic composition of the Nigerian Twitter population.
- A name matching dictionary to stimulate research in this area, and a modeling pipeline to compare the methods presented in this paper, publicly available at https://github.com/karimlasri/demographic_inference_nigeria/.

2 Background and Related Work

Prior work on predicting demographic characteristics of social media users produced methods for inferring various attributes such as gender (Fink, Kopecky, and Morawski 2012; Kim et al. 2017), age (Wang et al. 2019; Kim et al. 2017; Al Zamal, Liu, and Ruths 2021), ethnicity (Pennacchiotti and Popescu 2011; Preoţiuc-Pietro and Ungar 2018; Culotta, Kumar, and Cutler 2015), and occupation (Preoţiuc-Pietro, Lampos, and Aletras 2015; Pan et al. 2019). Most existing approaches used supervised learning, deriving features from user metadata, user-generated text and profile pictures (Fink, Kopecky, and Morawski 2012; Chakraborty et al. 2017; Wood-Doughty et al. 2018; Wang et al. 2019). Recent work has also shed light on the informational potential of network features (Li, Xu, and Lu 2015; Aletras and Chamberlain 2018; Pan et al. 2019) as well as data combination and transfer learning (Liu and Singh 2023). In the absence of labeled user data, researchers have resorted to rule-based methods as well as external data sources, such as county demographics (Mohammady and Culotta 2014), website traffic data (Culotta, Kumar, and Cutler 2015) and labeled name dictionaries (Mislove et al. 2011; Liu and Ruths 2013; Karimi et al. 2016). These methods have been applied at scale and provided valuable insights on the demographic composition of social media populations (Mislove et al. 2011; Mellon and Prosser 2017; Sloan et al. 2015). Yet, most of this work has focused on the Global North leaving little labeled data to perform demographic inference in the Global South. Consequently, we lack insight as to the demographic composition of social media populations in the Global South.

One of the few external data sources that are widely available in the Global South are lists of names associated with demographic characteristics, mostly originating from unofficial sources such as baby name websites. To perform demographic inference with this data source, the process consists in matching user names with labeled names and assigning the corresponding demographic characteristics to the matched user. Labeled name lists have long been used for demographic inference in several countries in the Global South, such as Nigeria (Rao et al. 2011) or more recently in India (Chaturvedi and Chaturvedi 2020) and Indonesia (Arafat et al. 2020). The main weakness of name matching is the limited coverage it offers, as not all user names are covered by name lists and not all users disclose their names on social media.

Graph propagation is a promising solution to expand the coverage offered by name matching. This class of methods consists of leveraging the social network structure and propagating information from labeled to unlabeled users. It relies on the *homophily* hypothesis, which states that users who share similar characteristics are more likely to be connected in a social network (McPherson, Smith-Lovin, and Cook 2001). Label Propagation (Wang and Zhang 2006) is a semi-supervised learning model that uses the graph structure to propagate label probabilities to unlabeled neighbors. This algorithm has been successfully used for social media user demographic inference in various contexts (Speriosu et al. 2011; Li, Xu, and Lu 2015). A more recent algorithm for graph-based propagation is the Graph Convolutional Neu-

Number of Nigerian Twitter users in our dataset	1.83M
Number of users with geolocation data	1.05M
Number of labeled names	2.5K
Number of users covered by name matching	1.08M
Number of annotated user profiles in our evaluation set	2K

Table 1: Summary statistics of the collected data

ral Network (GCN) (Kipf and Welling 2017) which has also successfully been used for demographic inference (Pan et al. 2019). Yet, despite existing work in a supervised learning context (Kim et al. 2017), it remains unclear how the most popular graph propagation models compare for large-scale demographic inference. To the best of our knowledge, our work is the first to provide an evaluation of graph-based propagation models for large-scale demographic inference in a low-resource context.

3 Data Description

3.1 Twitter Data

To collect our Twitter dataset, we first used the Twitter Decahose - a 10% random sample of all tweets globally - to identify approximately 1 million users with a profile location in Nigeria. Subsequently, we used the Twitter API to collect their timeline and the timeline of the users they mentioned. A snowball sampling over mentioned users was repeated four times. The first iteration yielded an additional 500,000 users, while the fourth revealed only 266 new users, indicating that the method enabled us to uncover all Nigerian users being mentioned by other users. In total, we gathered a dataset of about 1 billion tweets posted since January 2009 from the timelines of about 1.8 million users with a profile location in Nigeria. Recent estimates suggest that Nigeria has about 3 million Twitter users according to a recent report of the fact-check organization Africa Check (Okpi 2021). Additionally, Twitter documentation estimates that about 30-40% of tweets are posted by users with an identified profile location, which indicates that our dataset provides an excellent coverage for these users (Table 1.).

For all users, we also collect their their profile information as well as the list of their followees (the users they follow). User-level text data includes screen names, names, descriptions and all tweets, retweets and quoted tweets.

3.2 Name Dictionary

Following Rao et al. (2011), we collect a list of approximately 2,500 Nigerian first names from baby name websites distinguished by ethnicity and gender. Name lists are rare for minority ethnic groups; hence, we necessarily restrict our analysis to the four ethnic groups that represent the majority share (70%) of the Nigerian population, that is Hausa, Yoruba, Igbo and Fulani. Due to shared-religion naming conventions, Hausa and Fulani name lists are combined resulting in three distinct ethnic groups: Hausa-Fulani, Igbo and Yoruba. Because religion intersects significantly with ethnicity, we are able to derive religion labels from the predominance of Islam in the Hausa-Fulani community (95%) and Christianity in the Igbo community (98%)[1]. For the Yoruba community, which is 55% Muslim and 35% Christian, we reviewed and labeled first names based on local knowledge. For ambiguous cases where a name can be matched with many classes, as is the case with gender-neutral names, we assign scores based on their proportion when available[2]. Otherwise, we assign equal scores for possible classes.

3.3 Evaluation Set

To our knowledge, there is no publicly available labeled dataset for demographic inference on Nigerian Twitter. For evaluation, we randomly sample and label 2,000 Nigerian Twitter users for:

- *ethnicity* (Yoruba, Igbo or Hausa-Fulani)
- *religion* (Muslim or Christian)
- *gender* (Female or Male)

We choose to focus on majority sub-groups, for which there exists a wealth of resources. While this de facto excludes non-binary individuals, minority ethnic groups and traditional faiths, inference on these targeted groups is more likely to yield high accuracies and produce reliable predictions, as demonstrated by our name-matching strategy in §4.2. We further justify our choice in the Ethical Statement. Additionally, we tag accounts based on whether they belong to an organization or not to prefilter our test set at inference time.

The labeling was undertaken by a team of 4 Nigerian experts representing all major ethnic groups and with equal gender representation. Each user account was labeled by at least two annotators and disagreements were arbitrated by a third annotator. Cases with three labels and no majority label led to a team-wide discussion until an agreement was found. Our inference pipeline therefore relies on domain expertise and is specific to a population where names hold a strong signal for predicting the targeted demographic traits. Outside of this scope, this can be untrue and needs to be checked with experts at all times.

For ethnicity, gender and religion inference, we remove all users that do not have a label matching our target classes. Such cases correspond to accounts which were identified as belonging to an organization (9.8% of the test set), accounts for which the targeted attribute is hard to infer from the user's profile, or accounts that belong to a minority subgroup that we do not incorporate in our models. This results in a total of 1,599, 1,746 and 1,757 evaluation labels for ethnicity, religion and gender, respectively.

3.4 Descriptive Statistics

[1] https://www.familysearch.org/en/wiki/Nigeria_Religious_Records

[2] Gender proportions are available for the 1000 most popular Nigerian names at: https://forebears.io/nigeria/forenames

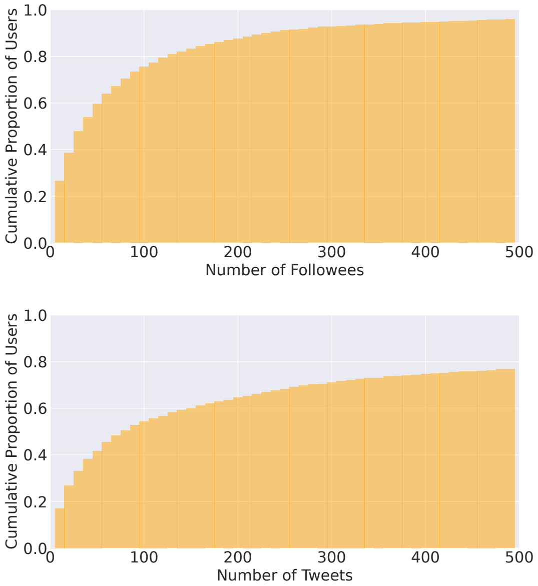


Figure 2: Cumulative proportion of users as a function of the number of connections (top) and the number of tweets (bottom)

User Activity Fig. 2, displays the cumulative proportion of users in our evaluation set as a function of their number of followees –that is the accounts they follow– and their number of tweets. We observe that close to 60% of users have fewer than 50 followees, with nearly 30% of users having fewer than 10 followees. This *a priori* makes our setting particularly challenging as we wish to rely on this connectivity information for all users without prefiltering. We also discover that more than 40% of users published less than 50 tweets. As we do not filter users based on their tweeting activity, this in principle should make text-based classification an arduous task. Indeed, previous studies impose a lower bound at 50 tweets to classify users based on the content they post (Fink, Kopecky, and Morawski 2012; Preoţiuc-Pietro, Lampos, and Aletras 2015).

Demographic Composition of Nigerian Twitter Based on findings from previous work (Okpi 2021) and the size of our user database, we estimate the share of Nigerians on Twitter to be approximately 1% of the general population. Next, we derive the demographic composition of the Nigerian Twitter population using the labeled random sample of users described earlier and present the results in Fig. 3.

Our first finding is that the Christian South is over-represented and the Muslim North is under-represented in the Nigerian Twitter population. Indeed, the share of Christian and Muslim users are respectively of 79.5% and 20.5% on Twitter when they amount to 45.9% and 53.5% in the Nigerian population. Together, the Yoruba and Igbo ethnic groups, mostly located respectively in the South-West and the South-East, amount to around 80% of the Nigerian Twitter population[3] and represent the double of their share in the general population. On the other hand, the inverse is true for the Hausa-Fulani ethnic group, mostly located in the North. Our second main finding is that men are over-represented on Twitter with 72% of Twitter users. Finally, we find that Twitter users are almost exclusively located in large urban areas, with 57% in Lagos alone, and with more than 80% located in four large Southern cities, namely Lagos, Abuja, Port Harcourt and Ibadan. Both the over-representation of men and populous areas are in line with the findings of Mislove et al.

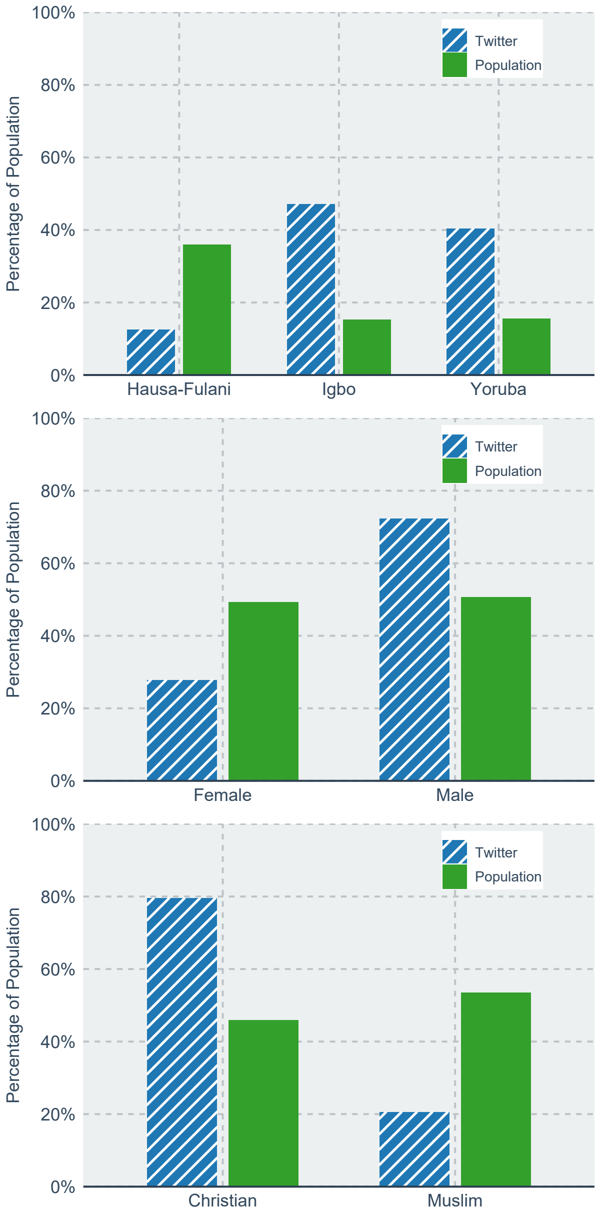


Figure 3: Share for each demographic attribute among Twitter users and in the Nigerian population

[3] Note that for ethnicity, around 10% of users in our random sample, not displayed here, belong to minority subgroups.

Attribute	Target Class	Prototypical Words
Ethnicity	Igbo	’ugwu.’, ’#biafra’, ’#freebiafra’, ’#nnamdikanu.’, ’#ipob’, ’onye’, ’ndi’, ’#peterobi4president2023.’, ’eze.’, ’ozo.’, ’#apeoplearecoming’
	Yoruba	’ekisola’, ”onalaja’s”, ’#shadesofus’, ’olorun’, ’#lra’, ”babatunde’s”, ’gbogbo’, ’temilolu’, ’#enikure!’, ’tó’, ’bá’, ’oro’, ’je’, ’ore’, ’ayodele’,’#femisolar’
	Hausa-Fulani	’daga’, ’wannan’, ’magana’, ’kyau’, ’samu’, ’kuma’, ’mutane’, ’ruwa’, ’allaah’, ’alhamdulillah’, ’bakin’, ’suratul’
Religion	Christian	’ernest’, ’jonah’, ’obi’, ’amen!!!’, ’onitsha’, ’enugu’, ’catholic’, ’harcourt’, ’priest’, ’testimonies’, ’#urchsalespoint’, ’#femisolar’, ’#rpn’, ’knowxup’
	Muslim	’daga’,’#quran’, ’wannan’, ’allaah’, ’hadith’, ’(saw)’, ’duniya’, ’ameen’, ’wallahi’, ’allah’, ’muhammad’, ’shuraim.’ ’#goforcompetence’, ’#fire’, ’babatunde’s’, ’saurari’, ’#naijatunez’, ’aragon’
Gender	Male	’shuraim.’, ’wlh’, ”sportmasta’s”, ”babatunde’s”, ’hospitality!’, ’#hustle’, ’dominance’, ’#euro2020’, ’#halamadrid’, ’#goforcompetence’, ’knowxup’, ’money’
	Female	’braid’, ’dresses’, ’shop:’, ’lace’, ’wigs’, ’glow’, ’ship’, ’slay’, ’makeup’, ’awww’, ’#etsy’, ’omg’, ’baby.’, ’dress’

Table 2: Examples of class-specific words picked from the 200 most prototypical tokens for each demographic attribute and each class. Class-specific words only represent a fraction of the most prototypical words and may convey negative stereotypes.

(2011) for the US Twitter population.

User Language As a sanity-check, we inspected the language makeup of tweets in our dataset. We found that 92% of tweets are categorized by the Twitter language algorithm as in English. When reading these tweets, we find that they are mostly in Nigerian Pidgin, an English-based creole language spoken as *lingua franca* across Nigeria. The very low prevalence of other Nigerian languages, such as Yoruba, Igbo or Hausa, is in line with the previous findings of a Nigerian Twitter population mostly based in southern large urban centers where Nigerian Pidgin is the first language.

4 Models

We start by comparing a variety of rule-based and supervised learning approaches to predict three demographic characteristics of Nigerian Twitter users: *ethnicity*, *religion* and *gender*. For supervised learning, we test the following features: unigram counts from user timelines, geolocation data and graph information from the followership network. Below, we describe the different modeling approaches evaluated.

4.1 Majority Baseline

We compute the majority baseline by computing the proportion of users which are part of the most common – or majority – class. This baseline can be trivially produced by a linear classifier which always returns the majority class, leaving a large margin for robust classification algorithms. The respective majority classes for ethnicity, religion and gender are Yoruba, Christian and Male.

4.2 Name-Matching Strategy

For each name N in our name dictionary, and each user U in our user database, we check whether N can be matched to either the screen name or the user name of U. Duplicates are dropped from the obtained list of matched names, yielding a list of n_m unique names. For each demographic attribute of interest, denoting n_c its number of target classes, we compute a score vector for each matched name based on our name dictionary $\mathbf{s}_i = (s_1, ..., s_{n_c})$, $1 \leq i \leq n_m$. Based on the intuition that longer matched names are more reliable to infer our target classes, we use the lengths of our matched names $l_1, ..., l_{n_m}$ to compute an averaged vector weighted by the matched names’ lengths,

$$\mathbf{s} = \frac{\sum_{1 \leq i \leq n_m} l_i \times \mathbf{s}_i}{\sum_{1 \leq i \leq n_m} l_i}$$

The highest dimension of this score vector yields the prediction $y = argmax(\mathbf{s})$.

	Ethnicity	*Gender*	*Religion*
Accuracy	0.84	0.93	0.85
Coverage	0.51	0.46	0.51
Acc. on whole	0.46	0.43	0.44

Table 3: Accuracy and coverage of our name matching method on our held-out test set. The accuracy on the whole dataset is obtained by multiplying the above two metrics.

As seen in Table 3, the name matching strategy yields robust performance on covered profiles. We further note that knowledge from this rule-based technique is imperfect, as we might not be capturing all ambiguous cases in our dictionary. Additionally, the matching procedure can lead to failures, especially when matching irrelevant short names which can accidentally appear in longer user names and pseudonyms. Most importantly, the matching only delivers partial coverage across classes, especially so for gender due to the non-negligible number of unisex names in our name database. We seek to address this issue by combining the matched names with graph-based propagation to infer attributes of other users.

4.3 Text-Based Classification

Our text-based approach builds on the knowledge collected from our name-matching approach. It consists of training models on users covered by the latter and testing them on a held-out test set of around 2,000 users.[4]

[4] Note that we filter out accounts which belong to organizations and minority groups, or for which it is difficult to assign a label, keeping a sample size of 1,6K to 1,8K users for each classification task, as mentioned in §3.3.

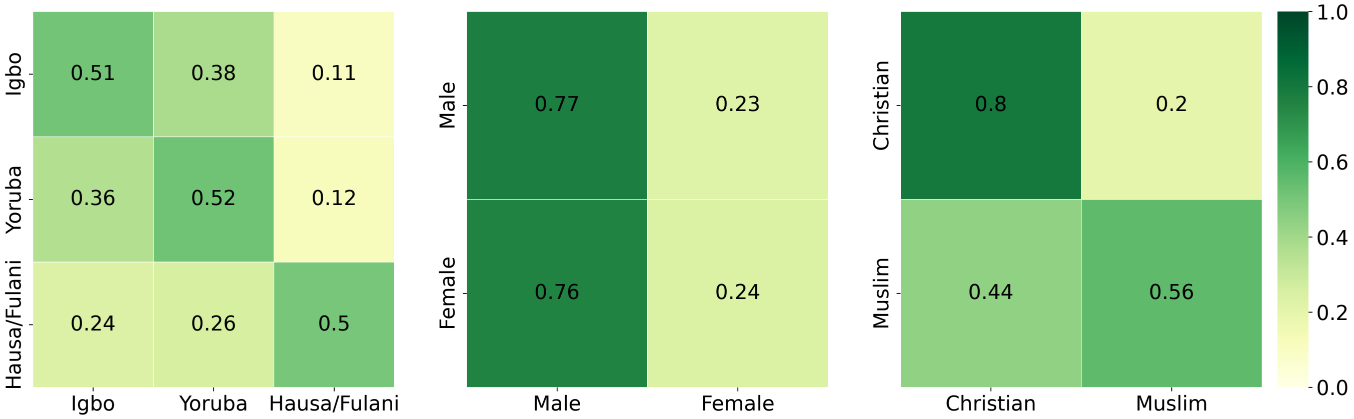


Figure 4: Homophily matrices for each user attribute. For users of a given class, each row represents the mean share of all classes among their connected users (followers and followees)

We resort to unigram-based count vectors rather than contextual representations from state-of-the-art pre-trained language models as currently there is no such model tailored to Nigerian Pidgin English, the main language in our tweets as discussed in §3.4. Collected tweets are converted into a unigram dataset which is further processed to create the following text-based features: simple occurrence counts for tokens, binary counts (i.e. a vector simply indicating the presence or absence of each token in a user's tweets) and tf-idf features. The vocabulary is filtered by removing stopwords and using a frequency threshold, i.e. we prefilter unigrams to keep only those which are used by at least 1,000 users. We then apply different feature selection techniques concurrently to select and retain the most relevant unigram features for each classification problem. As Table 3 shows, our name matching predictions are accurate overall, hence we use them as ground truth for feature selection. First, we try mutual information and chi-squared feature selection algorithms, selecting the top 200 features for each target class. We also employ a feature selection method described in previous work which yields prototypical words for each class (Pennacchiotti and Popescu 2011). For each token and each class (t, c), we compute a prototypicality score $s(t, c) = \frac{|t,c|}{|t|}$, that is the ratio between the frequency of t associated with class c and the frequency of t regardless of the output class. The top 200 most prototypical words for each class are selected to build the feature vector. We display a sample of such features in Table 2.

We find that users from each ethnic group disproportionately use words in the language of that ethnic group (e.g. 'ugwu', 'onye', 'ndi' for Igbo; 'enikure', 'tó', 'bá' for Yoruba; 'daga', 'wannan', 'magana' for Hausa-Fulani). First names specific to a given group are also disproportionately used by users from that group, both for ethnic (e.g Femisolar, Ekisola, Babatunde for Yoruba) and religious groups (e.g. Ernest, Jonah for Christians). Igbo users stand out by their focus on politics, with use of words related to the Biafra region, mostly populated by Igbos, as well as mentions of the Igbo presidential candidate Peter Obi and the Biafran activist Nnamdi Kanu. Finally, while Islam-related words are overrepresented in tweets posted by Muslims (e.g. 'quran', 'allah'), we find that Christians employ more words related to places where many Christians live (e.g. Onitsha, Anambra, Port Harcourt) or connected to Igbo people (e.g. Obi) who are in great majority Christian. Regarding gender, we find that some of the most prototypical tokens convey negative stereotypes, both for women (e.g. 'dresses', 'makeup') and men (e.g. 'dominance', and references to football). This exposes how harmful supervised text classification can be if features are not interrogated for fairness. This approach also picks up spurious tokens among the prototypical tokens which are not related nor specific to a group (e.g. 'text:@vanguardngrnews:', '#loveandships'). We also find that around 15% of users make use of prototypical tokens, which greatly limits their predictive power for the whole population.

In what follows, we use the notations *Unigram+Chi2*, *Unigram+MI*, and *Unigram+Proto* to respectively refer to the set of features obtained using chi-squared, mutual information and prototypical textual feature selection. For each set of features, we train three supervised models: a support vector machine classifier, a random forest classifier and a light gradient boosting machine classifier.

4.4 Label Propagation

The second model tested for demographic inference relies on graph-based information contained in the followership network. As the name matching approach allows us to cover 46% to 51% of users (see Table 3) depending on the attribute of interest, we leverage this information and the followership connections using a propagation method to induce the demographic classes of unlabeled users. Specifically, we first apply the method described in Wang and Zhang (2006), Linear Neighborhood Propagation, which we simply call Label

Propagation (LP) throughout this article. This algorithm departs from a set of labeled nodes and propagates this information to unlabeled users linearly from their neighborhood, that is users which they follow or which follow them. It assumes that each data point's label can be linearly retrieved from its neighborhood. For each demographic characteristic, we initialize a scores matrix $S^{(0)}$ in which we assign an initial score to each of our users, and for each of the n_c output classes. Denoting n_u the total number of users,

$$\forall i \leq n_u,\ c \leq n_c,\ \begin{cases} 0 \leq S^{(0)}_{i,c} \leq 1 \\ \sum_{c \leq n_c} S^{(0)}_{i,c} = 1 \end{cases}$$

To initialize this matrix, we use the name matching scores described in §4.2. We additionally initialize a weight matrix $W \in \mathcal{M}_{n_u \times n_u}(\{0,1\})$ as a binary matrix using followership information, where

$$W^b_{i,j} = \begin{cases} 1 & \text{if user } i \text{ follows user } j \\ 0 & \text{otherwise.} \end{cases}$$

We further compute a symmetric followership-based matrix $\hat{W} = W + W^\top$, assuming followership can be informative in both ways under our homophily assumption.[5] We finally normalize the rows of this matrix so that $\forall i \leq n_u, \sum_{j \leq n_u} \hat{W}_{i,j} = 1$.

At each iteration t, we update $S^{(t)}$ as follows:

$$\forall i \leq n_u, S^{(t+1)}_{i,c} = \alpha \hat{W} S^{(t)} + (1-\alpha) S^{(0)}$$

In practice, we use $\alpha = 0.5$. We repeat this procedure until convergence, at iteration t_f, and compute the prediction for each user i as $y_i = \text{argmax}(S^{(t_f)}_{i,\_})$.

This approach is appealing due to its simplicity. It builds on the homophily assumption (McPherson, Smith-Lovin, and Cook 2001), stating that users with demographic characteristics share edges in the followership network. As we depart from labels derived from our name-matching procedure, we report how homophilic users with names are with respect to our considered target attributes in Fig. 4. We observe that overall, users share their attributes with the majority of their connections, except for women. In this specific case, recall that our description of Twitter users displayed in Fig. 3 revealed a strong imbalance among users, with 72% of men among Nigerian users. The latter seems mirrored in our averaged homophily measures, even among women's connections. We further note that Muslims in turn do show strong homophily despite the strong imbalance between Muslims and Christians among users.

4.5 GCN with Multiple Features

Lastly, we make use of Graph Convolutional Networks (Kipf and Welling 2017). This model allows us to integrate both adjacency information and node-level information, combining features derived from user profiles with the followership network. By combining content-based and graph-based data, we expect to benefit from the advantages granted by both information sources. The model is initialized using the followership matrix introduced in §4.4, used for layer-wise propagation. The network is also fed with a node-level feature vector. Specifically, we concatenate several feature vectors representing the different information sources that we have access to. First, we input the final label propagation scores vector that were obtained in the previous section. We add user-level text features described in §4.3, and include geolocation information by projecting user geolocation into 34 localities. As seen in Table 1, we only have access to this information for 1.05M users, so we add an extra 'Unknown' locality and one-hot encode the resulting information into a 35-dimensional vector. Finally, we leverage information from the followership network to design vectors representing prototypical accounts for each class. To do so, we compute prototypicality scores using the same procedure used to select our textual prototypes, as described in §4.3. For each class, we keep the 200 most representative accounts based on these prototypicality scores, computed using the followership matrix, and append the resulting vectors to the features previously described.

	Ethnicity	*Gender*	*Religion*
Majority Baseline	0.47	0.72	0.76
Unigram + Chi2	0.47	0.72	0.75
Unigram + MI	0.47	0.71	0.77
Unigram + Proto	0.63	0.69	0.74
Label Propagation	**0.78**	0.79	**0.91**
GCN + User Features	0.76	**0.80**	0.90

Table 4: Accuracies on our held out test set for each of our different models

5 Results

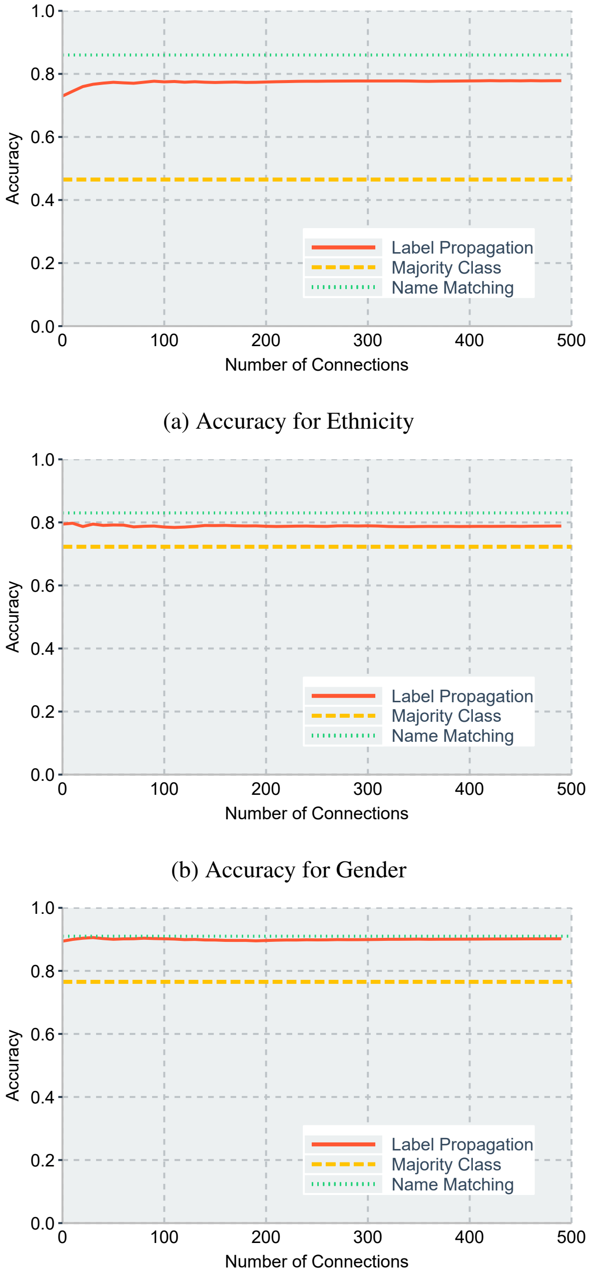
We report our classification results in Table 4. In addition to the accuracies of our various models, we display the majority baseline to represent the performance reached by a classifier which always predicts the most represented class among users.

5.1 Text-Based Classification

In Table 4, we display the performance of the best of our three classifier models using text-based features.[6] Models based on unigram data from users fail to outperform the majority baseline, achieving little to no improvement, on the contrary to previous studies (Fink, Kopecky, and Morawski 2012). This could be due to limited preprocessing of users based on their activity, as our goal is to infer demographic attributes for all users in our sample. Indeed, as displayed in Fig. 2, many of our users have limited posting activity. This could explain why our textual features are insufficiently predictive in our context despite the meaningfulness of some

[5] We first tested our models using only directed information, that is a followers-only and a followees-only matrix and our results were slightly better using the symmetric version.

[6] We do so for conciseness purposes, as the three classifiers introduced yield similar performance on each set of feature, close to the majority baseline.



(a) Accuracy for Ethnicity

(b) Accuracy for Gender

(c) Accuracy for Religion

Figure 5: Accuracy of the Label Propagation method as a function of the number of connections

of our prototypical words. We further note that prototypical feature selection allows us to gain performance on the ethnicity classification problem, but not for the two other classes despite extracting meaningful features. This indicates that limited posting activity contains ethnicity-specific content for a share of the users.

	Ethnicity	*Gender*	*Religion*
Accuracy	0.78	0.79	0.91
F1-score (weighted)	0.77	0.78	0.90

Table 5: Accuracy and weighted F1-score for the Label Propagation model

5.2 Followership-Based Label Propagation

The LP method coupled with our name matching procedure yields robust performance on all three targeted user attributes, thereby substantially outperforming the majority baseline. This finding is striking: in addition to being simple and interpretable, LP is robust and achieves competitive results.

As displayed in Fig. 5, the performance of this method is robust even for users with few connections. We note a slight decrease in performance for very low number of connections (≤ 30) when predicting ethnicity and religion, but the latter is small. While the accuracies reached are lower than those obtained using name matching, its wide coverage makes it more suited for our large-scale inference scenario as it can be applied to all users in our sample. Additionally, the algorithm's coverage and performance are stable for all targeted attributes after only two iterations. The wide coverage of our name matching procedure might be key to this quick convergence, along with connectivity properties of the followership graph which will require further investigation.

5.3 Graph Convolutional Network

While our GCN model integrates both followership information and several node-level features described in §4.5, we further note that it does not outperform LP. While textual features did not seem to improve performance above a majority baseline, this would not necessarily be true for other features included in our GCN model, such as geolocation. The fact that such features did not improve performance seems to indicate that the network relies mostly on our label propagation scores, while failing to leverage additional information found in the geolocation data of users. This in turn shows that the most crucial information for predicting our target classes arises from the homophily of users, captured by the connectivity of our extracted graphs. While only displaying one score in our figures and tables for this model type, we tried several combinations of input representations, by only keeping one or several of the features described in §4.5. The scores displayed in this paper are those resulting from the combination of all features, as removing some of them didn't sensibly change our results : the label propagation scores bear most, if not all of the information leveraged by our GCN.

To account for class imbalance, we additionally compute the weighted F1-score of our best-performing model LP in Table 5. Little difference between these two performance metrics shows that our model seems to equally handle all classes regardless of their respective shares, adding to its robustness in this large-scale, low-resource and imbalanced setting.

6 Discussion

Throughout this paper, we attempted to infer demographic attributes in a peculiar and challenging setting, as our objective is to make predictions for *every* user in our large database, regardless of their level of activity. Previous work focused mostly on inferring attributes on a carefully designed set, generally pre-filtered so that users bear sufficient information for supervised approaches to function. As an example, some previous studies only kept users with more than 50 tweets (Fink, Kopecky, and Morawski 2012; Preoţiuc-Pietro, Lampos, and Aletras 2015). In our scenario, it is rather desirable to elicit a method which is robust even for users with low activity, that is with a limited number of posts or connections. Our results shed light on (i) the limitation of text-based features in this scenario and (ii) the richness of followership information, even when the number of connections to other users is very limited (≤ 10) and for users who do not display their real name.

Despite using little data preprocessing, we are able to achieve performance that is competitive with previous work addressing our three target classes: ethnicity (Pennacchiotti and Popescu 2011; Culotta, Kumar, and Cutler 2015), gender (Al Zamal, Liu, and Ruths 2021; Fink, Kopecky, and Morawski 2012; Culotta, Kumar, and Cutler 2015; Mueller and Stumme 2016; Kim et al. 2017) and religion (Chaturvedi and Chaturvedi 2020). Hence, our combination of name-matching with label propagation can be applied in a large-scale real-world setting as long as a large share of users is covered by the rule-based matching, and the homophily hypothesis holds for the targeted attributes.

Our results should also draw awareness among users since making little information public, by not publishing much and not displaying their real name, does not prevent the inference of their attributes, as long as they provide reliable followership information to users sharing certain attributes.

7 Conclusion

Throughout this work, relying on a variety of features, we have compared different classification methods for large-scale inference of demographic attributes of social media users. We have shown that followership information can be leveraged to propagate targeted user attributes inferred using rule-based name-matching, a technique that requires little labeling efforts. The originality of our favored approach lies in solving our classification problems in a low-resource scenario. Indeed, we only annotate a set of 2,000 observations for testing purposes. We show that this technique performs well even for users with few followers and followees, making it appealing for inference at scale. Our ability to achieve high performance is a direct reflection of the homophily hypothesis, which holds for Nigerian Twitter users for our targeted attributes. Additionally, we demonstrate that text-based data-driven approaches not only pick up on undesirable stereotypical or spurious features, but also fail to provide key information overall to infer our targeted attributes of users with low activity. Lastly, we show that while integrating graph-based information and node-level features in a graph convolutional network bears promises, it does not significantly outperform label propagation technique, highlighting the effectiveness of this simple method in this context.

8 Future Work

Our work paves the way to a better understanding of conditions under which Label Propagation can be robustly deployed to infer attributes of social media users. While we covered a broad set of models and features, additional fine-grained analyses of network connectivity properties are required, to gain a deeper understanding of the amount of information required for propagation. For instance, future work could investigate how the initial amount of labeled users affects the algorithm's ability to achieve robust performance. Additionally, one could build on our observations that textual features are not well-suited when a large share of users have little published content. In particular, it would be interesting to investigate how different thresholds on users' posting activity affect text-based methods. Furthermore, the observation that data-driven prototypical unigram feature selection yields stereotypical associations calls for increased scrutiny into how such approaches bias text-based models in a way that discriminates unfairly. Another important avenue for future work would be to implement a model that is able to discriminate accounts that belong to organizations from those of users, based on information similar to that used in this paper. Finally, we note that we get fairly robust performance without having applied or built any bot detection model to our user database. More work is required to understand the extend to which discriminating bot accounts beforehand could affect the results of our demographic inference pipeline.

Ethical Statement

Addressing the previously unknown, we describe Nigerian Twitter and highlight that interpretable and fair algorithms can provide comparably high performance to more advanced but less transparent and potentially more biased methods. Broadly speaking, beyond describing Nigerian Twitter, we expect that the utility of this approach will be evident in future work that relies on demographic inference to evaluate policy impact. Yet, we acknowledge that our approach, as well as broader research aimed at inferring demographic attributes of users, may raise several ethical concerns. For instance, the tools developed can be used for profiling purposes in pursuit of malicious objectives. Due to the accessibility of available tools and the high risk of re-identification (Rocher, Hendrickx, and De Montjoye 2019), the data used for development and evaluation may be sensitive and require confidentiality.

Moreover, we are aware that name-based demographic inference may disproportionately miscategorize minority groups and individuals which can have serious empirical and ethical consequences, as extensively discussed in Lockhart, King, and Munsch (2023). In this work, we only provide a modeling pipeline aimed at accurately inferring demographic traits of social media users and thus remain agnostic on the usage of such tool. In doing so, we leave it to practitioners and researchers who wish to incorporate such information in their analysis to estimate whether their work is ethically desirable. It is our responsibility, however, to draw attention to a number of critical points and limitations, which should be taken into account when evaluating whether future work building on our methods is ethically justified. In the following, we build on the suggestions formulated by Lockhart, King, and Munsch (2023).

First, our method produces attributes based on external ascriptions, and therefore should be used in case studies where one is interested in external ascription, e.g. how social media users perceive each other, instead of focusing on a user's true sense of self-identity. Additionally, our method builds on labels produced and revised by local domain experts who have a strong knowledge of the extent to which names or profiles signal certain traits, as described in §3.3. Labels assigned without such precaution might not be as accurate and produce erroneous inferences. We try to limit subjective judgements and individual biases in the annotation process by duplicating the labeling task among our four experts and by making sure disagreements are arbitrated. This however does not impede unfair biases being shared by all of our annotators, despite their level of expertise. Further, for each demographic attribute, we limit ourselves to traits which can be retrieved from a user's name with high accuracy in a given population, as demonstrated by our name matching results. In doing so, we fail to capture a variety of subgroups, including non-binary individuals, ethnic minorities, and traditional faiths, as resources are lacking for these groups which could result in developing poorly performing models. This draws a limitation of our approach: while providing accurate predictions at the aggregated level, it disregards minority subgroups, which limits inclusivity. Any downstream usage of methods or data similar to ours should take this limitation into account. Finally, as our pipeline is accurate at an aggregated level, downstream applications relying on similar data should preferably make use of demographic predictions at the group level, as individual predictions might contain erroneous associations which could add confounds to the modeling pipeline.

Acknowledgments

We thank Arinze Alex Okoyeugha, Fatima Sulaiman and Khashau Adenike Eleburuike for their excellent research assistance. This work was supported by funding from the United Kingdom's Foreign Commonwealth and Development Office (FCDO) and the World Bank's Research Support Budget. This work was also supported in part through the NYU IT High Performance Computing resources, services, and staff expertise.

The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations or those of the Executive Directors of the World Bank or the governments they represent.

References

Al Zamal, F.; Liu, W.; and Ruths, D. 2021. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(1): 387–390.

Aletras, N.; and Chamberlain, B. P. 2018. Predicting Twitter User Socioeconomic Attributes with Network and Language Information. In *Proceedings of the 29th on Hypertext and Social Media*, HT '18, 20–24. New York, NY, USA: Association for Computing Machinery. ISBN 9781450354271.

Arafat, T. A.; Budi, I.; Mahendra, R.; and Salehah, D. A. 2020. Demographic analysis of candidates supporter in twitter during indonesian presidential election 2019. In *2020 International Conference on ICT for Smart Society (ICISS)*, 1–6. IEEE.

Chakraborty, A.; Messias, J.; Benevenuto, F.; Ghosh, S.; Ganguly, N.; and Gummadi, K. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, 22–31.

Chaturvedi, R.; and Chaturvedi, S. 2020. It's All in the Name: A Character Based Approach To Infer Religion. *arXiv preprint arXiv:2010.14479*.

Culotta, A.; Kumar, N. R.; and Cutler, J. 2015. Predicting the demographics of twitter users from website traffic data. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Fink, C.; Kopecky, J.; and Morawski, M. 2012. Inferring gender from the content of tweets: A region specific example. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 6, 459–462.

Gao, H.; Wang, Z.; and Ji, S. 2018. Large-Scale Learnable Graph Convolutional Networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, 1416–1424. New York, NY, USA: Association for Computing Machinery. ISBN 9781450355520.

Karimi, F.; Wagner, C.; Lemmerich, F.; Jadidi, M.; and Strohmaier, M. 2016. Inferring gender from names on the web: A comparative evaluation of gender detection methods. In *Proceedings of the 25th International conference companion on World Wide Web*, 53–54.

Kim, S. M.; Xu, Q.; Qu, L.; Wan, S.; and Paris, C. 2017. Demographic Inference on Twitter using Recursive Neural Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 471–477. Vancouver, Canada: Association for Computational Linguistics.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Kryvasheyeu, Y.; Chen, H.; Obradovich, N.; Moro, E.; Van Hentenryck, P.; Fowler, J.; and Cebrian, M. 2016. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3): e1500779.

Li, P.; Xu, J.; and Lu, T.-C. 2015. Leveraging Homophily to Infer Demographic Attributes: Inferring the Age of Twitter Users Using Label Propagation. In *Proceedings of Workshop on Information In Networks (WIN15)*.

Liu, W.; and Ruths, D. 2013. What's in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*.

Liu, Y.; and Singh, L. 2023. Combining vs. Transferring Knowledge: Investigating Strategies for Improving Demographic Inference in Low Resource Settings. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, 868–876.

Llorente, A.; Garcia-Herranz, M.; Cebrian, M.; and Moro, E. 2015. Social media fingerprints of unemployment. *PloS one*, 10(5): e0128692.

Lockhart, J. W.; King, M. M.; and Munsch, C. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour*, 1–12.

Lorenz-Spreen, P.; Oswald, L.; Lewandowsky, S.; and Hertwig, R. 2023. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature human behaviour*, 7(1): 74–101.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1): 415–444.

Mellon, J.; and Prosser, C. 2017. Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3): 2053168017720008.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 554–557.

Mohammady, E.; and Culotta, A. 2014. Using County Demographics to Infer Attributes of Twitter Users. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 7–16. Baltimore, Maryland: Association for Computational Linguistics.

Mueller, J.; and Stumme, G. 2016. Gender inference using statistical name characteristics in twitter. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, 1–8.

Okpi, A. 2021. https://africacheck.org/fact-checks/reports/forty-million-twitter-users-nigeria-how-pollsters-flawed-figure-became-fact. Accessed: 2023-01-14.

Palotti, J.; Adler, N.; Morales-Guzman, A.; Villaveces, J.; Sekara, V.; Garcia Herranz, M.; Al-Asad, M.; and Weber, I. 2020. Monitoring of the Venezuelan exodus through Facebook's advertising platform. *Plos one*, 15(2): e0229175.

Pan, J.; Bhardwaj, R.; Lu, W.; Chieu, H. L.; Pan, X.; and Puay, N. Y. 2019. Twitter Homophily: Network Based Prediction of User's Occupation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2633–2638. Florence, Italy: Association for Computational Linguistics.

Pennacchiotti, M.; and Popescu, A.-M. 2011. A Machine Learning Approach to Twitter User Classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1): 281–288.

Preoţiuc-Pietro, D.; Lampos, V.; and Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1754–1764. Beijing, China: Association for Computational Linguistics.

Preoţiuc-Pietro, D.; and Ungar, L. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1534–1545. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Rao, D.; Paul, M.; Fink, C.; Yarowsky, D.; Oates, T.; and Coppersmith, G. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, 598–601.

Rocher, L.; Hendrickx, J. M.; and De Montjoye, Y.-A. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature communications*, 10(1): 1–9.

Sloan, L.; Morgan, J.; Burnap, P.; and 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.

Speriosu, M.; Sudan, N.; Upadhyay, S.; and Baldridge, J. 2011. Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, 53–63. Edinburgh, Scotland: Association for Computational Linguistics.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Eighth international AAAI conference on weblogs and social media*.

Wang, F.; and Zhang, C. 2006. Label propagation through linear neighborhoods. In *Proceedings of the 23rd international conference on Machine learning*, 985–992.

Wang, Z.; Hale, S.; Adelani, D. I.; Grabowicz, P.; Hartman, T.; Flöck, F.; and Jurgens, D. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference*, 2056–2067.

Wood-Doughty, Z.; Andrews, N.; Marvin, R.; and Dredze, M. 2018. Predicting Twitter User Demographics from Names Alone. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 105–111. New Orleans, Louisiana, USA: Association for Computational Linguistics.