

Demographic Inference and Representative Population Estimates from Multilingual Social Media Data

Zijian Wang*
Stanford University
zijwang@stanford.edu

Scott A. Hale
University of Oxford &
Alan Turing Institute
scott.hale@oii.ox.ac.uk

David Adelani
MPI-SWS & Saarland
University
dadelani@mpi-sws.org

Przemyslaw A. Grabowicz
MPI-SWS & University of
Massachusetts
grabowicz@cs.umass.edu

Timo Hartmann
GESIS Cologne
timo.hartmann@gesis.org

Fabian Flöck
GESIS Cologne
fabian.floeck@gesis.org

David Jurgens[†]
University of Michigan
jurgens@umich.edu

ABSTRACT

Social media provide access to behavioural data at an unprecedented scale and granularity. However, using these data to understand phenomena in a broader population is difficult due to their non-representativeness and the bias of statistical inference tools towards dominant languages and groups. While demographic attribute inference could be used to mitigate such bias, current techniques are almost entirely monolingual and fail to work in a global environment. We address these challenges by combining multilingual demographic inference with post-stratification to create a more representative population sample. To learn demographic attributes, we create a new multimodal deep neural architecture for joint classification of age, gender, and organization-status of social media users that operates in 32 languages. This method substantially outperforms current state of the art while also reducing algorithmic bias. To correct for sampling biases, we propose fully interpretable multilevel regression methods that estimate inclusion probabilities from inferred joint population counts and ground-truth population counts.

In a large experiment over multilingual heterogeneous European regions, we show that our demographic inference and bias correction together allow for more accurate estimates of populations and make a significant step towards representative social sensing in downstream applications with multilingual social media.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Computing methodologies** → **Natural language processing**; **Object identification**; **Neural networks**.

KEYWORDS

Demographics; Post-stratification; Social Media; Latent Attribute Inference; Inclusion Probabilities; Multilingual; Deep Learning

*Work performed while at the University of Michigan

[†]Corresponding senior author

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.

ACM ISBN 978-1-4503-6674-8/19/05.

<https://doi.org/10.1145/3308558.3313684>

ACM Reference Format:

Zijian Wang, Scott A. Hale, David Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens. 2019. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*, May 13–17, 2019, San Francisco, CA, USA. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3308558.3313684>

1 INTRODUCTION

Data representing the attitudes and behaviors of a (national) base population is of great importance to policy making, social science research and commercial prediction tasks, but representative surveys are expensive and infrequent. Social media data has been proposed as a real-time and inexpensive way to measure social phenomena—this area of research is known as “social sensing.” Social media data has been used, with various levels of success, in areas such as infectious disease [30, 31, 51], migration and tourism [7, 50, 81], and box office takings for films [21, 55].

However, despite social media data being easily accessed, large in scale, and detailed, it is generally not representative of indicator variables measured in any broader (offline) population due to various biases [43, 67]. Especially the self-selection of users joining and participating on any given social media platform is a large source of error. In the UK for instance, Twitter users are more likely to be male and young compared to the national population [71], and in the U.S. men and residents of densely populated areas are overrepresented [35], while there exists a mix of over- and undersampling on Twitter for users with specific racial backgrounds [56]. In general, however, the exact inclusion probabilities for a specific demographic group on any platform in a given country or region are unknown. And while not all social sensing tasks require representative data [3], recent studies have pointed out the fallacies in predicting phenomena via social sensing without controlling for sampling biases in social media data [28, 29, 44], with research applications aiming to draw inferences for nation-wide target populations often relying on such attribute data. Consequently, a systematic methodology for estimating them is needed.

Survey analysis researchers have dealt with non-representative polls and non-response through sample re-weighting—with post-stratification being a well-known technique [10, 37]. These methods make use of basic demographics attributes: most prominently age, gender and location [60, 77]. Where demographic details are provided, post-stratification has been a valuable technique [77, 81].

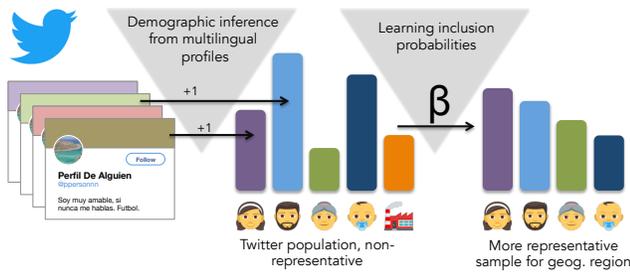


Figure 1: Our two-stage approach debiases non-representative Twitter data by (i) inferring demographics with state-of-the-art accuracy from multilingual, profile-only data and (ii) learning inclusion probabilities to create more representative samples for Europe-wide subregions

However, most social media platforms do not provide demographic data about their users, making it difficult to know or correct for such biases. Although promising techniques for extracting demographic attributes from social media have been proposed, there exists no robust multilingual approach for this task. We address these challenges by inferring basic demographic attributes and correcting for selection biases on a large sample of multilingual profile data from Twitter (Figure 1), one of the most used platforms for social sensing given its ease for obtaining data.

This work offers the following three contributions. First, we introduce a new multilingual, multimodal, multi-attribute deep learning system for inferring demographics of users (§2). This pipeline is built from the ground up to enable inference in 32 different languages, addressing the need for such methods beyond English and outperforming state-of-the-art methods on the tasks of predicting the age, gender and is-organization state of Twitter users. We show this system reduces algorithmic bias with respect to skin tone over the best-performing commercial systems. Second, we formalize a statistical framework for models that debias non-representative samples (§3). Our models explicitly learn per-stratum inclusion probabilities from data, whereas typical post-stratification methods either assume their knowledge or focus on obtaining a post-stratified estimator of a response variable. Third, in real-world evaluations, we show that our debiasing models significantly reduce the error rate in comparison to a model without post-stratification on the inferred demographic attributes (§4). In these ways, we make a significant step towards *representative social sensing* in downstream applications using multilingual social media.

2 DEMOGRAPHIC INFERENCE

We propose a new demographic inference model for three attributes: gender, age, and a binary organization indicator (“is-organization”). The first two attributes, gender and age, are widely reported in census data and are core features to accurately measure demographic biases. The third attribute was selected based on a known confounder for people-based studies of online platforms: the presence of accounts belonging to organizations [2, 54], which shall be distinguished from individual accounts. Next, we describe the model, its data, and its evaluation.

2.1 Classification Task

We consider the task of estimating population count in Twitter given a stream of user tweets. This scenario is motivated by the common research use case for social sensing on Twitter, where a researcher consumes one of the Twitter streams (e.g., the 10% or 1% sample streams) and desires to make some analysis from the tweets of a broader phenomena that extends beyond Twitter. In such a scenario, many users are expected to appear rarely—possibly only once—which hinders the use of existing demographic methods that rely upon large amounts of text to infer gender [13, 17, 18, 25, 45, 49, 72, 73, 82] and age [33, 58, 59, 63, 65, 68].

Instead, we avoid making inferences based on tweets and focus on using only the information associated with a user’s account, which allows our method to easily scale to large volumes of users without the need for significant quantities of text. Not stratifying users based on tweets has the added advantage of preventing downstream demographic biases to any social sensing tasks that analyze the same language. Ultimately, our method uses four sources of information, username, screen name, biography, and profile image, and is designed to operate effectively on multilingual data.

We formalize the three classification tasks as follows. Gender and is-organization are modeled as binary classification tasks.¹ Age is known to be a difficult inference task in social media [58, 82]. Here, we opt to categorize age in four levels: ≤ 18 , (18, 30), [30, 40), [40, 99). Our choice is motivated by (i) our ability to align these age ranges with census data reports and surveys [74, 75], and (ii) the difficulty of the task (even for humans) when finer-grained age classes are used [23, 76], while (iii) still making the resulting classes amenable to downstream tasks. While other attributes are possible for use in debiasing (e.g., education, income, etc.), ground truth population statistics for these attributes are not widely available in censuses, making it difficult to train and evaluate any model.

2.2 Method and Training Procedure

Demographic attributes are often expressed in both image and text. We propose a new multimodal inference model with a novel training technique that leverages both of these sources to significantly augment our predictive accuracy. We first describe the model and each modality and then discuss the training procedure.

2.2.1 Model. The full model comprises two separate pipelines for processing a profile image and each of the three text sources of information, as shown in Figure 2. Here, we use a shared pipeline for all three attributes, using multi-task learning for the output. We refer to this as the **M3** model, after its multimodal, multilingual, and multi-attribute abilities. Following, we describe the architectures of each modality and how they are combined.

Image Model The image classifier was constructed using DenseNet [38] based on initial performance experiments over current state-of-the-art vision models on our data. We scale all profile images to 224x224 to meet the input size of the model.

¹ We base our notion of gender as one of performance [22], in which individuals adapt their style, name, and picture to (de)emphasize certain aspects of their gender identity [24]. While our training data uses self-declared binary gender identities, the M3 model reports gender identity as a continuum using prediction probability along a graded scale. We recognize that the current approach does not yet support non-binary gender identities, which we view as an important future task for full representation.

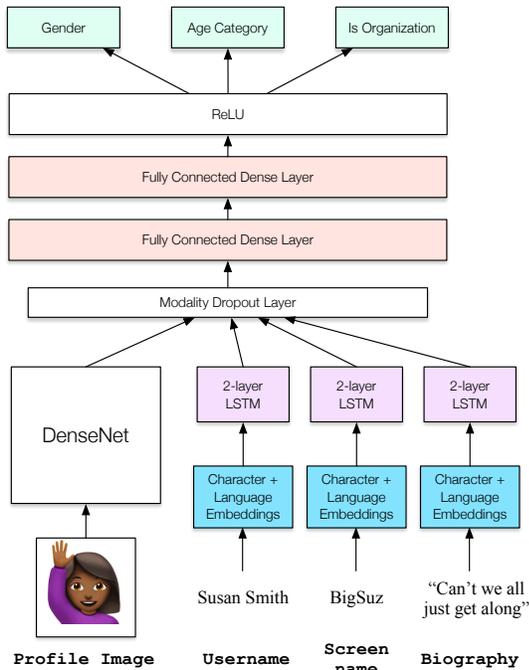


Figure 2: The M3 model for inferring gender, age, and organization-identity from image and text data.

Text Model Separate character-based neural models were trained for each text input and then a multi-input model was tuned based on the weights from each single model. Embeddings for username and biography were limited to the 3000 most frequent characters in our training corpus, with remaining characters represented with an encoding for their unicode category; screen name are embedded in the ASCII range. We use 2-stack bidirectional character-level Long-short Term Memory (LSTM) architecture [36] to capture necessary information. To incorporate potential different meanings in different languages for the same character, we concatenate a separate trainable language embedding to each character embedding and add a fully connected layer before being fed into an LSTM. The fusion of character-based input with language embeddings provides two critical benefits that let us scale to the linguistic diversity seen in global platforms like twitter. (1) Character-based RNNs are able to capture morphological regularity within a language and substantially reduce the parameter space due to the reduced vocabulary size of the model (i.e., the set of characters) as compared with word-based models with orders-of-magnitude larger vocabularies [15, 27, 46]. (2) By fusing language embeddings with character embeddings, the joint embedding lets us represent shared linguistic regularity *across* languages, which lets us better scale to the multilingual environment.

Full Multimodal Model To construct the full model, separate models are fully trained for image and text for all attributes. Then, the softmax output layer of each is removed and the two models are joined into a new model with a modality drop-out layer (described next), two fully connected layers of size 2048, a Rectified Linear Unit (ReLU) layer [57], and separate output layers for each task. This

full model is then fine-tuned using the pretrained single-modality models’ parameters for initialization.

2.2.2 Training Procedure. Rather than training M3 end-to-end initially, we combine three procedures for the training process: co-training, multilingual data augmentation, and modality drop out. *Co-Training* Both image and text can have clear signal of demographics. Given the potentially-limited training data for speakers of less common languages, we take advantage of the multiple views of the same users by incorporating co-training: a semi-supervised learning technique. Co-training first learns a separate classifier for each view using any labeled examples. The most confident predictions of each classifier on the unlabeled data are then used to iteratively construct additional labeled training data [11]. Ardehaly and Culotta [4] recently showed co-training was effective for demographic classification in social media data using images and text in a graph-based co-training learning procedure.

In comparison to the variance of gender and age expressions in textual data across sociolinguistic contexts [6, 70], image data provides a more universal depiction, acting as an analog to a universal pivot language in machine translation [80]. Thus, we train the image-based portion of the M3 model first, which identifies high confidence labeled data from an unlabeled multilingual dataset. Second, the text-based portion of the M3 model is trained using the original dataset plus the text data associated with any high-confidence image-based classifications. Finally, we train the full M3 model with the training dataset and all high-confidence predictions from the unlabeled data. All models are trained as multi-task, where a common input representation and architecture is used to predict each output tasks. The probabilities for each task were represented using a top softmax layer.

Multilingual Data Augmentation The M3 model is designed to operate in online environments on tens of languages. However, few approaches and little labeled data exists for this task for languages other than English, with some exceptions [16, 32, 62]. Even with co-training, the relative differences in number of speakers between languages (e.g., English vs. Slovenian) create a substantial class imbalance which would prevent the model from effectively learning signals of each attribute in users’ biographies written in less-frequent languages. Therefore, we perform data augmentation by using automatic machine translation of the training data.

Given the scale of our training data (described later in §2.3), the use of online translation services is prohibitive both in terms of time and cost. Instead, we opt to use word-based translation using the MUSE bilingual dictionaries [20]. For each instance in the training data, all words in the bilingual dictionary are replaced with the translation. A translated instance is only kept if 80% or more of its words could be translated. While word substitution typically produces translations that are more incorrect and less fluent, we observed that the remaining biography translations were often of reasonable quality due to the declarative nature of biographies, which often consist primarily of lists of readily-translatable noun descriptions or attributes of the user without using full grammatical sentences. Such translations also largely capture topical and semantic regularity, even if the morphology is not precise.

Modality Drop-out The co-training model uses a conjoining architecture with pre-trained weights from the image model and the

multi-input text model. To facilitate the model using each source of information, we introduce a new technique, modality dropout, that can fully hide a source of information (e.g., the biography) during training. We add an extra input dropout with a probability of 0.25 (i.e., keep 3 out of 4 input sources on average) to better regularize the model and support use cases where not all 4 input sources are available. When doing dropout, images were replaced by a random matrix while text was turned to a specific empty embedding.

2.3 Data

Training and evaluating M3 uses five distinct datasets: (1) a large dataset of Twitter profiles whose gender and age are heuristically identified, predominantly in English, (2) a curated dataset of Twitter accounts belonging to organizations, (3) an image dataset of faces from IMDB and Wikipedia, (4) a massive unlabeled set of users, and (5) a crowdsourced dataset for all three attributes that spans all 32 languages in our study. We describe each next.

2.3.1 Heuristically-identified users. Heuristically-identified users were drawn from a 10% sample of Twitter from 2014 to 2017 on Twitter, where roughly 40% of the users are English speakers [34]. Here, we identify gender and age from fixed expression in users’ biographies that signal age or gender, e.g., “mother of two wonderful kids” or “26 y/o dude” using a gender- or age-signaling word. Age data was further augmented by identifying tweets wishing another user a happy birthday (cf. Al Zamal et al. [1]). These pattern-based approaches are known to be highly precise for certain attributes when properly constructed [8, 9]. Patterns were created for five languages: English, Spanish, French, German, and Swedish.² Organization accounts were manually curated by identifying 676 Twitter lists that primarily contained organizations, such as non-profits, local clubs, companies, or municipal services. These lists resulted in 59.92K unique organization accounts. For heuristically-identified users and organization accounts, we collect their current profile image, screen name, username, and biography. User age is adjusted to the present day. For heuristically-identified users, the biography is altered to replace the gender- or age-indicating word with a special token so that the model is forced to look for additional cues to recognize each attribute. The IMDB-WIKI dataset [66] consists of 523,051 images from headshots of actors in the IMDB and profile pictures in Wikipedia. This dataset provides an auxiliary source of information for fine-tuning the image-based part of the M3 model. These three datasets constitute our *initial data* prior to performing the three-step annotation procedure. Table 1 shows the detailed statistics of the data for each step.

Co-training depends on having access to a large set of unlabeled users whose high-confidence labels from one modality (e.g., profile image) can be used to augment the training set for another modality. Here, we collect 36.97M profiles speaking one of the 32 languages commonly spoken in Europe but for which we have no ground truth label. Languages were identified using CLD2 [53]. Note that users in this unlabeled dataset will only become training instances if one view’s classifier (text or image) labels them with high confidence.

2.3.2 Crowdsourcing Data. Current gender, age, and organization datasets are nearly all produced for English-speaking users, which

²Languages were selected to capture linguistic diversity and where high precision demographic-identify patterns could be created easily.

Table 1: Sizes of datasets used for training and testing. Incremental numbers are shown for the outputs of data-augmentation from co-training and translation.

	Dataset	Gender	Age	Is-Org.
	<i>Initial</i>	3.98M	1.20M	59.92K
	After Image Co-Training	+2.26M	+0.62M	+6.31M
	After Translation Augmentation	+5.74M	+1.54M	+4.74M
	After Text Co-Training	+8.28M	+0.78M	+11.87M
	<i>Final</i>	14.53M	2.61M	23.86M
	Held-out Evaluation	0.38M	0.36M	6.91K

prevents us from evaluating the M3 model in the multilingual environment. Therefore, we constructed a new dataset of up to 200 randomly-selected Twitter users speaking one of each of the 32 languages in our study. Three annotators on Figure Eight³ were shown the usernames, screen names, short biographies, and profile images of each Twitter account. We instructed workers to determine gender and age of a given user from two drop-down menus based only on this information. The selection options regarding gender also included a category for profiles belonging to multiple persons, and to a non-person (e.g., organization/bot). Regarding age, annotators were given seven categories to choose from: ≤ 18 , (18, 30), [30, 40), [40, 50), [50, 60), [60, 70), ≥ 69 . They could select “don’t know” for all questions, if they felt not enough information was available. We employed Figure Eight’s selection mechanism for speakers of the target language of the set to be annotated, or chose workers by their country of residence, if one of that country’s main spoken languages was identical with the target language of the task. Additionally, each job included a set of 30 test questions, annotated as a gold standard by the authors. These were used to screen for spam but also for any non-speakers of the target language remaining after our first filter; to this end, at least 10 questions were only solvable by sufficiently understanding the target language.

Annotation reliability was measured for each Figure Eight job using Krippendorff’s α [48]. For the computations of α , we excluded all English test-questions from the 31 non-English jobs, profiles with only one annotation, and included “don’t know” answers as an answer option. To establish a measure of how well workers did on the age classification in respect to the age brackets classified by M3, we collapsed age bracket encodings over 40 into one category. Across all languages, the mean α for the gender annotations lies at 0.81 and at 0.57 regarding the age brackets, hinting at the high difficulty of the Twitter profile age classification task for humans. For a separate measure, we recoded all gender and organization annotations into the two categories “is organization” and “is not organization” and calculated Krippendorff’s α based on these data, resulting in a mean of approximately 0.75 across all languages, showing that workers could identify non-personal accounts reasonably well. The final dataset is constructed from all instances where at least two annotators agreed on the label for a particular attribute (majority vote), resulting in 4,732 profiles across 32 languages.

2.3.3 Data Partitions. All self-reported labeled data is partitioned into 80% train, 10% development, and 10% test splits. No official

³<https://www.figure-eight.com/>

train/test/dev splits are available for the IMDB-WIKI datasets so we perform our own partition with the same percentages as for self-reported. During model training, only the development set for self-reporting users are used for model selection. Owing to its small size, crowdsourcing data was only used for testing purposes. For the organizational data, we create a full dataset with non-organizational accounts by randomly sampling accounts from the heuristically labeled data to attain a 1:9 ratio, following the report in McCorrison et al. [54] that organizations make up 10% of the accounts on Twitter.

2.3.4 Co-Training Setup. The training data is balanced for each stage, where we oversampled gender and organization status, and undersampled the two most-frequent classes (0-18 and 19-29) while oversampling the other two age classes. High confidence thresholds were set to 0.9 for gender and organization status in both the image and text modalities. As a four-class classification task, the high-confidence for age is set to 0.7 in the image classifier, based on early examination of the prediction quality. We observed that age is more readily predictable in text and therefore used a threshold of 0.9 to keep quality high while also providing a sufficient number of new accounts for co-training.

Separate image and text models are each trained for 10 epochs. Due to the imbalanced number of instances per attribute (Table 1), each epoch contains 10000 steps where each step contains a mini-batch of 128 instances balanced across the three attributes. This setup ensures the model sees instances of the smaller datasets (age and organization). The image model is trained first, and its high-confidence instances are then used to train the text model. The final co-training model was trained for 5 epochs with input dropout and another 3 epochs without input dropout for better convergence. Amsgrad optimizers [64] were used in all training process, with a learning rate of 0.001 for ground-up models and 0.0005 for fine-tuning models. Text models were trained on one GPU, while the image and co-training models were trained in parallel on 4 GPUs. The full pipeline was built in PyTorch [61].

2.4 Evaluations

Accurate inference of representative attitudes from social media relies upon accurate stratification approaches. Here, we test the M3 model against current state of the art systems to describe the potential for error to affect demographic bias.

2.4.1 Comparison Systems and Data. The M3 model is compared with state-of-the-art systems from each modality and attribute. For images, we compare with Face++ [42] and Microsoft Face API [5] on age and gender performance. For text, we compare with three current state-of-art systems for inferring gender from usernames, genderperformr [78], demographer [47] and Jaech et al. [39], which could feasibly operate on names from multiple languages.

No multilingual organization recognition system exists; so, we limit our evaluation to the only publicly-available dataset of person-vs-organization which was scored by the Humanizr [54] and Demographer [79] systems. This data consists of a uniform sample across Twitter accounts, which were approximately 10% organizations. Here, we recollect the user profiles for their 20,273 accounts, of which 18,587 (91.6%) were still available as of October 2018,

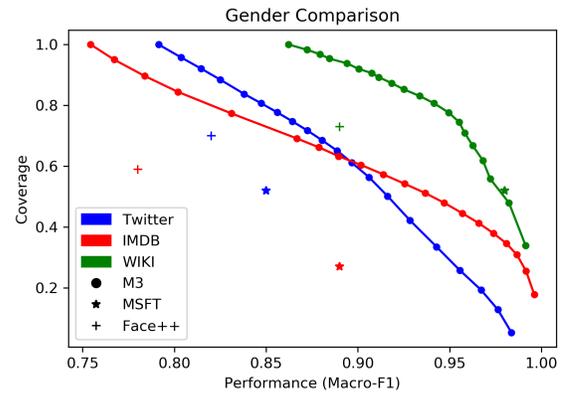


Figure 3: Model performances on gender classification for the three image datasets, as measured through F1. Coverage equals % of returned results at varying confidence levels for M3 and is the maximum available for the commercial APIs.

which matched the distribution originally reported. Their model was evaluated on the dataset using cross-validation, whereas we treat the entire data as a test set and report performance.

2.4.2 Gender Recognition. M3 produces state-of-the-art gender recognition for each of the three image-based datasets (IMDB, WIKI, and Twitter), as seen in Figure 3. M3 provides significant improvements in performance (F1) at each comparison system’s level of recall.⁴ The one exception to this trend is the performance of the MSFT classifier on the WIKI dataset, which has highly similar performance to M3. Given the MSFT’s model large jump in performance, we hypothesize that their model may have been trained on Wikipedia data, although this cannot be confirmed. We observed that the increase in coverage of M3 is due in part to the model learning non-facial attributes associated with more than one gender, e.g., certain sports with the male gender, which allow M3 to process profile photos that other models cannot.

On the text-based data, M3 outperforms all systems except one when using just the username and masking all other information, shown in Table 2. When the model is allowed to see all other text information, M3 performance improves substantially, indicating it can effectively fuse several sources of information. While this latter setup uses more information than comparison systems, in practice, the Twitter API includes all text information used by M3 by default; so the only additional step is downloading the profile image.

2.4.3 Age Recognition. Age recognition is a significantly more difficult task in social media, as seen by the performances in Figure 4. The M3 model offers similar performance to commercial models on the IMDB and WIKI datasets, which primarily feature face-forward headshots in good lighting. However, in real-world Twitter profiles (shown in blue), the M3 model substantially outperforms both in F1 at each of their respective coverage levels, with 0.16 and 0.11 absolute improvements over Face++ and Microsoft, respectively.

⁴The coverage reported is the maximum available from the commercial APIs, which did not find a face or did not provide a gender/age estimate for each photo.

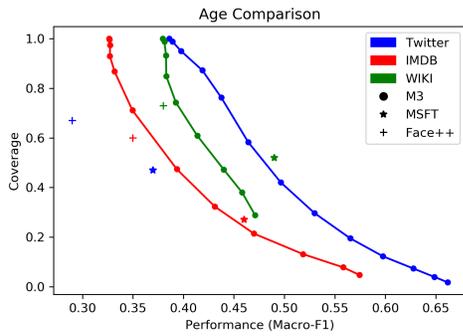


Figure 4: Model performances on age classification for three datasets, as measured through F1 on classified images and Coverage on the % of images classified.

2.4.4 *Organization Recognition.* For any task that draws inferences about human actors, non-human and specifically organizational accounts are a major source of error. The results of Table 3 show that our model is able to identify these accounts with 16.3% higher accuracy than the next closest system with no drop in performance at recognizing humans. Results on our test set from our organization data (§2.3) in Table 4 (top right) show similar performance, with an overall F1 of 0.898 that indicates high performance for both classes.

2.4.5 *Multilingual Performance.* As a full test of M3, we evaluate it against our crowdsourced account labels in 32 languages, shown in Table 4 top right. Performance in this multilingual setting is on par with the performance on the primarily-English heuristically-labeled data, though age classification remains the most difficult task. Figure 5 shows the performance per language for each attribute. Performance at predicting gender and is-organization are similar for most languages, while age performance varies significantly by language from 0.28 F1 for Bosnian to 0.73 for Slovenian and Welsh. These results together indicate that M3 is sufficiently accurate in the multilingual European environment.

Table 2: Performance comparison on gender using text methods. †Demographer only has a coverage of 0.974.

Method	Macro-F1
M3: Full Text	0.907
M3: Username Only	0.828
GenderPerformr [78]	0.835
Demographer [47]	0.781 [†]
Jaech and Ostendorf [39]	0.763

Table 3: Performance at recognizing organizational accounts, measured as accuracy per class, following [54]

Method	Person	Organization
M3	0.986	0.807
Demographer [79]	0.973	0.644
Humanizr [54]	0.982	0.586

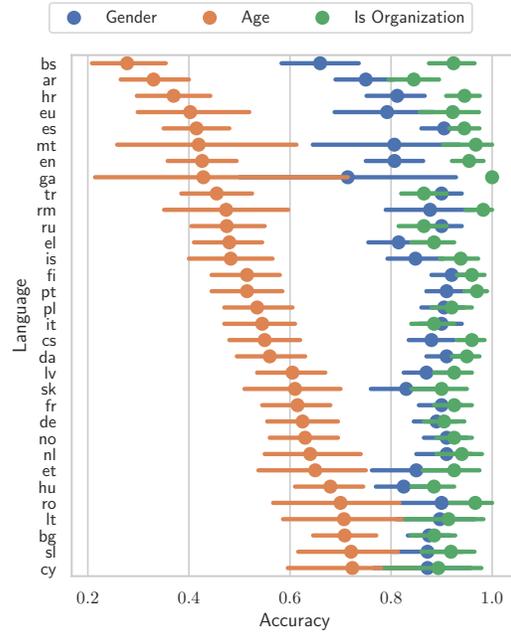


Figure 5: M3 performance on the multilingual dataset in 32 languages. Bars show 95% confidence intervals of accuracy.

2.4.6 *Ablation Study.* To examine which parts of the system are contributing to performance, we perform an ablation study by (i) restricting the model to one modality and (ii) training a model without using co-training or translation for data augmentation. We test each model on (i) the heuristically-labeled and organisation data and (ii) the multilingual data, which is representative of the performance on the data used in our downstream task.

The results shown in Table 4 reveal two main trends. First, the model benefits from both modalities. The removal of textual information causes the biggest performance drop. Second, although co-training and translation hurt performance in the heuristically-labeled data, they produced a substantially better model when evaluated on multilingual data. Since the heuristically-labeled data is primarily in English, translation potentially adds noise and forces the model to represent a larger space of inputs that are not represented in the test data. This performance difference indicates that the two data augmentation techniques are highly beneficial when bootstrapping a model from mostly monolingual data.

Table 4: Performance (Macro-F1) for (1) full M3 model, (2-3) model with modality dropouts, (4) model trained without co-training or translation, (5-6) Random and Majority-class baselines.

M3 Model	Heuristically-labeled			Multilingual		
	Gender	Age	Org-Status	Gender	Age	Org-Status
Full	0.918	0.522	0.898	0.915	0.425	0.898
w/ Text Dropout	0.743	0.349	0.837	0.837	0.359	0.892
w/ Image Dropout	0.905	0.493	0.823	0.862	0.372	0.822
w/o Co-T. & Trans.	0.918	0.449	0.917	0.875	0.325	0.660
Random	0.494	0.181	0.500	0.494	0.211	0.466
Majority	0.377	0.214	0.500	0.334	0.187	0.210

Table 5: Performance on the Gender Shades dataset (cf. Table 4 in [12]). Gender classification performance as measured by the positive predictive value (PPV), true positive rate (TPR), and false positive rate (FPR).

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	100.0
	TPR	93.7	96.5	91.7	87.1	99.3	92.1	83.7	100.0	98.7
	FPR	6.3	8.3	3.5	12.9	0.7	16.3	7.9	1.3	0.0
Face++	PPV	90.0	78.7	99.3	83.5	95.3	65.5	99.3	94.0	99.2
	TPR	90.0	98.9	85.1	83.5	95.3	98.8	76.6	98.9	92.9
	FPR	10.0	14.9	1.1	16.5	4.7	23.4	1.2	7.1	1.1
M3	PPV	96.7	99.4	94.5	94.0	99.3	99.6	89.3	99.3	99.2
	TPR	96.5	92.8	99.6	93.4	99.3	86.0	99.7	99.0	99.5
	FPR	4.2	0.4	7.2	7.7	0.8	0.3	14.0	0.5	1.0

2.4.7 Test for Algorithmic Bias. Image-based models for gender recognition are known to suffer from algorithmic bias in recognizing darker-skinned individuals [12]. This bias is thought to be a result of non-representative training data the underrepresentation of darker skin tones in the training data. As our focus is on European countries, such a bias could be present in the M3 data. However, because our co-training procedure uses a large unlabeled set of users from across the globe, these users can potentially provide a more representative sample and reduce algorithmic disparity. To test for algorithmic bias, we compare the M3 model’s performance on the Gender Shades dataset [12], which contains 3964 gender-annotated facial images balanced across light and dark skin tones. In this earlier algorithmic audit, the Microsoft (MSFT) and Face++ gender classifiers performed substantially worse on women, especially darker skinned women.

The gender inference performance, shown in Table 5, reveals that M3 has substantially less algorithmic bias relative to the two best-performing commercial systems also tested on Gender Shades. Our model significantly improves performance on dark-skinned women in comparison to the other two systems which had the largest performance disparity on that demographic. However, M3 is least accurate on darker-skinned males, which indicates that additional work is needed to reach performance parity in skin tones. Nevertheless, this analysis also indicates that M3 is more suitable than existing systems for operating in the global environment and increases the robustness of downstream social sensing applications.

3 LEARNING INCLUSION PROBABILITIES

In social sensing studies, the results of measurements performed on a given platform, e.g., a social media site, are studied to understand the behavior of a population. Often, such measurements are biased [67], as individuals with certain demographics are more likely to join these platforms, e.g., young people may be more likely to join Twitter. Obtaining representative estimates in these scenarios is challenging, as the probability of an individual with given demographics to be on a given platform, also referred to as *inclusion probability* in sampling methodology, is typically unknown.

Here, we estimate these inclusion probabilities as a function of demographics. To this end, we learn the debiasing coefficients on the grounds of statistical survey analysis with missing data [10, 52, 69]. Specifically, we derive estimators for the numbers of individuals by making an assumption that per-strata inclusion probabilities are

equal for individuals within a group (e.g., a country) and different between groups (e.g., between countries). This assumption allows us to get both across-groups global debiasing coefficients (i.e., the biases of Twitter users) and the group-specific debiasing coefficients (i.e., per-country demographics on Twitter, e.g., developed countries may have more and diverse users on Twitter). Our approach learns the inclusion probabilities, while typical post-stratification methods either assume these probabilities are known or focus on obtaining a post-stratified estimator of a response variable.

3.1 Formulation of Debiasing Models

Consider a population U of $N = |U|$ individuals with certain demographics. For simplicity, we focus on the case of only two discrete demographic variables, say age a and gender g , which are distributed following $P_N(a, g)$, but the following reasoning applies to other demographics as well. Out of these N individuals, say, $M \leq N$ joined a certain online platform with a probability depending on their demographics. For instance, individuals joined Twitter and younger people were more likely to join than older ones.

In survey analysis, the probability that an individual with certain demographics joined a certain platform corresponds to the inclusion probability of the stratum representing these demographics [10, 69]. In the simplest scenario, when the probability of joining the platform is homogeneous in time and across individuals, this probability can be expressed as the ratio of the number of Twitter users with the demographics a and g to all individuals, i.e., $\pi(a, g) = \frac{M(a, g)}{N(a, g)} = \frac{MP_M(a, g)}{NP_N(a, g)}$, where $P_M(a, g)$ and $P_N(a, g)$ are the distributions of demographics of Twitter users and the overall population, respectively. However, the inclusion probability may vary between individuals. To account for this, we discuss the homogeneity of inclusion probability for a given partition of the population. A partition of the population U is defined through non-overlapping and non-empty subsets U_i of the population that together constitute U , i.e., $\bigcup_i U_i = U$. Typically such subsets will have a certain meaning, e.g., a natural partition of a population is a split by countries, regions, or cities. The total number of individuals in the subset i is N^i and the number of individuals having particular demographics is $N^i(a, g)$.

3.1.1 Homogeneous bias. The inclusion probability, which governs the bias, is homogeneous with respect to a given set of demographics and a given partition of the population, if and only if $\pi^i(a, g) = \pi(a, g)$ for each subset i of the partition, i.e., the inclusion probability does not depend on the elements of the partition.

If the inclusion probability is homogeneous, then we can write

$$N^i = \sum_a \sum_g N^i(a, g) = \sum_a \sum_g \frac{M^i(a, g)}{\pi(a, g)}. \quad (1)$$

Thus, to obtain the inclusion probabilities, we regress N^i against $M^i(a, g)$. If the condition of homogeneity holds, then the regression coefficients are equal to $1/\pi(a, g)$.

3.1.2 Inhomogeneous bias. If the inclusion probability is inhomogeneous for given demographics and partition, then we shall model each subset of the partition separately or use a different model that relaxes the homogeneity assumption by specifying the functional form of inhomogeneity. Here, we consider the inhomogeneity of

the form

$$\pi^i(a, g) = M^i(a, g)^v f_1(a) f_2(g), \quad (2)$$

where v is an unknown exponent and f_1 and f_2 are unknown functions.⁵ Under this assumption, $\log N^i(a, g) = \log \frac{M^i(a, g)}{\pi^i(a, g)} = (1 - v) \log M^i(a, g) - \log f_1(a) - \log f_2(g)$. This time, to obtain the debiasing coefficients and corresponding inclusion probabilities, we regress the ground-truth population, $\log N^i(a, g)$, against the biased measurements, $\log M^i(a, g)$, and the demographic indicator variables, i.e.,

$$\log N^i(a, g) = \beta_1 \log M^i(a, g) + \sum_{\bar{a}} \beta_{\bar{a}} \delta_{\bar{a}a} + \sum_{\bar{g}} \beta_{\bar{g}} \delta_{\bar{g}g} \quad (3)$$

where δ_{ij} is Kronecker delta, i.e., $\delta_{ij} = 0$ if $i \neq j$ and $\delta_{ij} = 1$ if $i = j$, $\beta_1 = (1 - v)$, $\beta_a = -\log f_1(a)$, and $\beta_g = -\log f_2(g)$. From these regression coefficients we can obtain the inclusion probability $\pi^i(a, g)$ of a set of samples i with the demographics a and g via Equation 2. This debiasing model has been proposed recently by Zagheni et al. to predict the number of migrants based on Facebook data [81], but its derivation and explicit formal interpretation have not been provided until now, to the best of our knowledge. Note that this estimation method of inclusion probabilities requires the ground-truth joint counts $N^i(a, g)$ at the time of training, whereas the model based on the homogeneity assumption requires only the total counts N^i . The availability of the joint counts is often limited in practice because of insufficient number of samples per stratum.

3.2 Discussion

We derived debiasing models by making homogeneity or inhomogeneity assumptions. These models predict population size, so they can be evaluated by measuring the error of predictions in cross-validation settings or via model selection methods. In this way, we test which assumption is closer to reality and find the most accurate method for the given dataset. We evaluate these models in the next section using the population of regions in EU countries.

4 EUROPEAN POPULATION INFERENCE FROM TWITTER DATA

Here, we use the debiasing models formulated in the previous section to obtain the debiasing coefficients and the corresponding inclusion probabilities for Twitter users of different countries. To this end, we regress the country-level ground-truth number of people living in a certain location against the number of Twitter users of different demographics.

4.1 Debiasing Models for Population Inference

More specifically, we evaluate five models requiring different amounts of data. The first model is a baseline, the next three models are based on the assumption of homogeneous inclusion probabilities (Equation 1), whereas the last one is based on the inhomogeneity assumption (Equation 3):

$N \sim M$ is our base model that uses only the total population count from the census (N) and Twitter (M).

$N \sim \sum_g M(g)$ uses gender marginal counts only (i.e., the total counts of males and females not broken down by ages).

$N \sim \sum_a M(a)$ uses age marginal counts only.

$N \sim \sum_{a, g} M(a, g)$ uses the joint histograms inferred from Twitter but only the total population values from the census.

$\log N(a, g) \sim \log M(a, g) + a + g$ uses the joint histograms inferred from Twitter and the joint histograms from the census.

Note that Twitter users are biased in various ways: the platform is more accessible to tech-savvy individuals and citizens of various countries use it to a different extent [34]. To distinguish the global effect of Twitter on the overall bias from the local effect of country, we use multilevel models. Namely, all slopes and intercepts in the introduced models have random effects specific to a country. Thus, from fixed effects of the model, we obtain global debiasing coefficients (i.e., the biases of Twitter users) and, from random effects, the country-specific debiasing coefficients (i.e., a given country has its own bias towards Twitter).

Note that the homogeneity and inhomogeneity assumptions apply within each country separately. Namely, we group samples by regions of a country, described next. Then, the homogeneity assumption translates into the same inclusion probability for a given demographic across all regions of a country, whereas the inhomogeneity assumption translates into a dependence of the inclusion probability on regions of a country that follows Equation 2.

4.2 Data

We retrieved joint population distributions for age and gender on a regional level (NUTS3) for 26 countries of the European Union plus four EFTA members as made available through the CensusHub system of Eurostat, the European Statistical Office.⁶ All census data is from 2011, the most recent year for which comprehensive census data was collectively reported to Eurostat. NUTS3 regions are the finest-grained level of the ‘‘Nomenclature of Territorial Units for Statistics’’ used as a standard for statistical reporting in EU member states [26]. They are based on existing local administrative boundaries and are usually at the level of local districts. Although sizes differ to a certain extent per country, NUTS3 provide the most standardized cross-country geographical units to subdivide populations. We use the 2010 iteration of the NUTS3 regions as these correspond with the 2011 census subdivisions and is the most recent census available.

Our sample of social media users is derived from the random 10% stream of tweets from Twitter. We recorded all users observed from September 2015 to January 2016 and downloaded profile photos for all users in spring 2018. The location of each user is inferred using the method of Compton et al. [19]. This model predicts a user’s latitude and longitude and was shown to be least biased with respect to urban and rural areas [41], which is important given the diversity in population centers in our study. The model was trained on a social network of 781M edges constructed from Twitter data spanning 2012 to 2017. Five-fold cross validation reports a median inference error of 7.9km, which is sufficiently accurate for the geographic granularity we use here. Each user is assigned to a NUTS3 region in Europe if the inferred latitude-longitude pair of

⁵Note that the aforementioned homogeneity assumption is a special case of this assumption, i.e., this assumption relaxes the homogeneity assumption.

⁶<https://ec.europa.eu/CensusHub2/>: Luxembourg and Belgium were not available. Additional EFTA countries: Switzerland, Iceland, Liechtenstein and Norway.

that user lies within the boundaries of the NUTS3 regions; users not in these regions are discarded. Ultimately, we obtain a dataset of 3,202,964 users within the NUTS3 regions for our study. In the remainder, we will refer to NUTS3 regions simply as “regions.”

The age, gender, and is-organization variables for each user in our dataset are inferred using M3. As an additional experiment to quantify the impact of including non-human accounts, we ignore any organization classification and treat all accounts as humans, grouping them according to their inferred demographics.

4.3 Results

We evaluate the debiasing models in the following cross-validation settings: leave one region out, leave one country out (i.e., leave out all regions from a given country), and leave one stratum out (e.g., leave out only males aged 30-39). Per case, we measure the mean absolute percentage error of the population estimates for the left-out samples, i.e., $MAPE(N) = \frac{100\%}{n} \sum_i \frac{|\hat{N}_i - N_i|}{N_i}$, where \hat{N}_i and N_i are the predicted and actual population sizes, respectively, and the sum is over all regions. Note that the model $\log N(a, g) \sim \log M(a, g) + a + g$ operates naturally in the log space; hence, before calculating the error of this model we first exponentiate the predicted population sizes to make the results more comparable across all models. Its results should further be compared to the remaining models with care, as the effects of moving to log space cannot be clearly untangled from gains due to learning on joint age and gender attributes.

The results of the leave-one-region-out evaluation show a clear benefit to debiasing based on the inferred demographics (Figure 6). For the $N \sim M$ model without demographics, MAPE is 88%. The inclusion of the inferred gender or age demographics in the debiasing models, via $N \sim \sum M(a)$ and $N \sim \sum M(g)$, decreases MAPE to 59% and 61%, respectively. Including inferred joint distributions, e.g., the number of males of age 30-39, in the model $N \sim \sum M(a, g)$, decreases MAPE further to 54%, even though we do not use joint-distribution data from the census for training. In fact, the use of census joint distributions at the stage of model training along with a move to log space, via $\log N(a, g) \sim \log M(a, g) + a + g$, further improves the accuracy, bringing the MAPE down to 33%. To understand the effect of accounts belonging to organizations on this population prediction task, we compare the results of the debiasing models with and without removing the organizations (Figure 6). Removing organizations results in a small reduction in error for all but the model trained with only marginal gender counts, though these differences are not statistically significant. A further analysis shows that the presence of organizations is heavily skewed towards populous metropolitan cities and most regions have very few organizations (i.e., organizations are not equally geographically distributed). However, for populous regions, organizations can create significant error.

To gain further insights into the introduced models, we show scatter plots of true and predicted population sizes for each model (Figure 7). The model with joint Twitter and census counts is noticeably closer to the $y = x$ line, likely because this model is trained in log space. For this model, we also depict the geographical variation of MAPE across the EU regions (Figure 8).

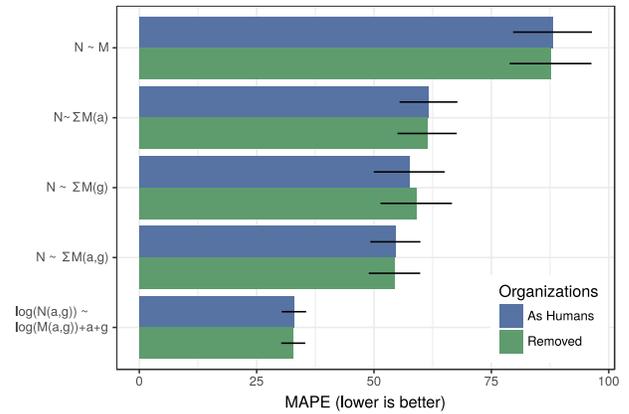


Figure 6: Performance on leave-one-region-out population inference across different debiasing models. Bars show mean MAPE(N) with 95% confidence intervals. The results of the last model are exponentiated to compute MAPE(N).

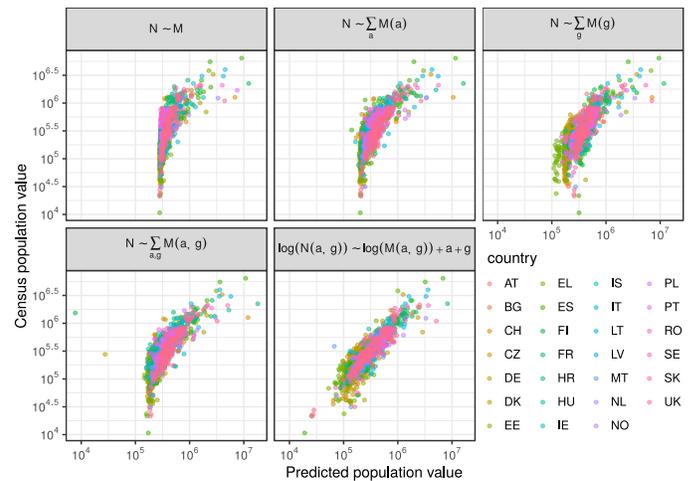


Figure 7: Comparing the census population and debiased estimates for leave-one-region-out evaluations shows that the model with joint Twitter and census distributions correlates best with the real values for large and small regions.

The above results test predictions when one region is left out and its population is predicted. However, in different circumstances, more or less census information could be available. We test two other cross-validation settings: leave one stratum out (e.g., leave out only females aged 30-39) and leave one country out (i.e., leave out all regions from a given country). The latter reflects the generalizability of the model to completely unseen countries where platform adoption probabilities and country-specific biases are not known. The evaluation results for all three cross-validation settings are plotted in Figure 9 for the most accurate model, i.e., $\log N(a, g) \sim \log M(a, g) + a + g$. From the first additional cross-validation setting, we learn that hiding the population sizes for a specific stratum in all regions results in a minimal penalty to the prediction accuracy. For the second additional cross-validation setting, we see that the error rate nearly doubles with an average MAPE of 81%. This result suggests that knowledge of country-specific

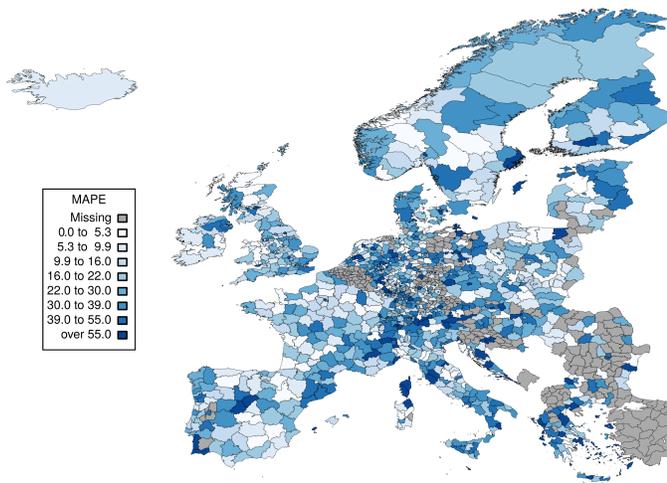


Figure 8: MAPE for $\log N(a, g) \sim \log M(a, g) + a + g$ model across regions shows that many regions have population estimate errors below 10% of the true population, with high-error regions found infrequently in all countries. For greyed regions we either miss NUT3 data or Twitter counts are zero.

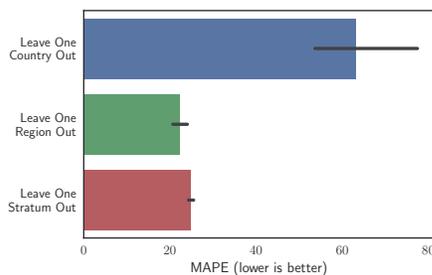


Figure 9: Comparisons of the model with joint Twitter and census distributions as increasing amounts of information are held out from training: (i) leaving out one stratum, (ii) leaving out all the strata from one region, and (iii) leaving out all regions in one country.

platform biases is important for accurate estimates and that at least some regions within the country should be seen during a model’s training time to reach high performance.

4.4 Discussion of Debiasing Results and Potential Sources of Error

Debiasing social media data samples is a difficult task, but our results show that automatic post-stratification with respect to inferred age and gender notably improves population estimation results. Furthermore, our predictive models are fully interpretable, which allows us to estimate the inclusion probabilities and share them with the research community for future reuse.

Even when joint distributions are not given by the census, inferring joint distributions from social media data with the model $N \sim M(a, g)$ provides a significant increase in accuracy in population prediction tasks compared to the baseline without debiasing.

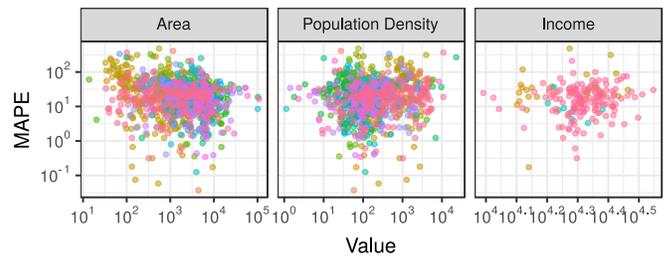


Figure 10: Comparing the model with joint Twitter and census distributions MAPE to region area (km^2), population density ($\text{people}/\text{km}^2$), and income (USD per capita) shows that the model performance is not biased towards particular types of regions. Colors represent countries as in Figure 7.

However, the model $\log N(a, g) \sim \log M(a, g) + a + g$ is notably more accurate, suggesting that the assumption of homogeneity does not hold for the partition of citizens of a particular country, gender, and age into regions. Note, however, that the predictions of the most accurate model are not perfect. This may be caused by the time mismatch of about 5 years between the Twitter and NUTS3 datasets. On the other hand, this point opens the door for developing debiasing models based on other inhomogeneity assumptions and searching better partitions, e.g., our multilevel model could have more levels to capture biases shared by smaller regions within a country.

To learn more about potential confounders, we compare the MAPE of the model $\log N(a, g) \sim \log M(a, g) + a + g$ to three variables: the area of the region, its population density, and average income (Figure 10). NUTS3 regions are defined by each country and vary considerably in land area from city-states in Germany to large regions in Sweden (up to 98,000 km^2). We do not find correlations of land area with the MAPE error. Hecht and Stephens [35] found a clear urban–rural bias in social media data: there tend to be more users, more information, and higher quality information per capita within metropolitan areas. Despite this bias, we find our estimates not heavily correlated with population density, suggesting our models do well at debiasing these differences in inclusion probabilities. Only two countries (Czech Republic and Norway) show a correlation that is statically different from zero across our models. Finally, income has been suggested as an important explanatory variable for who uses Internet-based platforms and hence could be a potential confound to our estimates [14, 40]. However, we similarly find that income and MAPE are not correlated significantly in our models.

5 CONCLUSION

The everyday opinions expressed in social media provide promising opportunities for measuring population-level statistics for health metrics, political outcomes, or general attitudes. However, social media is a non-representative sample of the population due to demographic skew in usage frequencies and access rates. As such, any direct estimate from a platform is likely biased towards certain demographics. This work provides a holistic solution to this problem by developing a novel method for assigning users to demographic strata and exploiting it in a regression framework for debiasing that allows direct estimation of the probability of an individual

with given demographics to be on the given social media platform. Our work provides three main contributions. First, we introduce a state-of-the-art neural system for multi-attribute classification in 32 languages. This contribution also includes the system release and the creation of a new dataset of gender, age, and is-organization annotations in 32 languages. Second, we derive a series of models to debias social media measurements and provide their explicit formal interpretations. Third, in a massive study of all of Europe, we show that our two methods are able to infer regional population counts accurately and provide demographic corrections for all downstream measurements on the grounds of the estimated inclusion probabilities. These results pave the way for more accurate social sensing by laying a foundation of representative population sampling in social media. Code, software, and debiasing coefficients pertaining to this work are released for public use at <https://github.com/euagendas/>.

ACKNOWLEDGEMENTS

This research has received funding through the Volkswagen Foundation.

REFERENCES

- [1] Faiyaz Al Zamil, Wendy Liu, and Derek Ruths. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *Proceedings of ICWSM*.
- [2] Sultan Alzahrani, Chinmay Gore, Amin Salehi, and Hasan Davulcu. 2018. Finding Organizational Accounts Based on Structural and Behavioral Factors on Twitter. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. Springer, 164–175.
- [3] Jisun An and Ingmar Weber. 2015. Whom should we sense in “social sensing”—Analyzing which users work best for social media now-casting. *EPJ Data Science* 4, 1 (30 Nov 2015), 22. <https://doi.org/10.1140/epjds/s13688-015-0058-9>
- [4] Ehsan Mohammady Ardehalay and Aron Culotta. 2017. Co-training for Demographic Classification Using Deep Learning from Label Proportions. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 1017–1024.
- [5] Microsoft Azure. 2018. Cognitive Services. <https://azure.microsoft.com/en-us/services/cognitive-services/>
- [6] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18, 2 (2014), 135–160.
- [7] Daniele Barchiesi, Helen Susannah Moat, Christian Alis, Steven Bishop, and Tobias Preis. 2015. Quantifying International Travel Flows Using Flickr. *PLOS ONE* 10, 7 (07 2015), 1–8. <https://doi.org/10.1371/journal.pone.0128470>
- [8] Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I’m a believer: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 181–186.
- [9] Shane Bergsma and Benjamin Van Durme. 2013. Using Conceptual Class Attributes to Characterize Social Media Users. In *Proc. ACL*.
- [10] Jelke G Bethlehem and Wouter J Keller. 1987. Linear weighting of sample survey data. *Journal of Official Statistics* 3, 2 (1987), 141–153.
- [11] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. ACM, 92–100.
- [12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 77–91.
- [13] Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A Comparative Study of Demographic Attribute Inference in Twitter. In *Proceedings of ICWSM*, Vol. 15. 590–593.
- [14] Menzie D Chinn and Robert W Fairlie. 2007. The determinants of the global digital divide: a cross-country analysis of computer and internet penetration. *Oxford Economic Papers* 59, 1 (2007), 16–44.
- [15] Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1693–1703.
- [16] Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1136–1145.
- [17] Jennifer Coates. 1998. *Language and gender: A reader*. Wiley-Blackwell.
- [18] Jennifer Coates. 2015. *Women, men and language: A sociolinguistic account of gender differences in language*. Routledge.
- [19] Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. In *IEEE Conference on BigData*.
- [20] Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087* (2017).
- [21] Brian de Silva and Ryan Compton. 2014. Prediction of Foreign Box Office Revenues Based on Wikipedia Page Activity. *CoRR* abs/1405.5924 (2014). arXiv:1405.5924 <http://arxiv.org/abs/1405.5924>
- [22] Victoria Pruijn DeFrancisco, Catherine Helen Palczewski, and Danielle D McGeough. 2013. *Gender in communication: A critical introduction*. Sage Publications.
- [23] Hedwige Dehon and Serge Brédart. 2001. An ‘other-race’ effect in age estimation from faces. *Perception* 30, 9 (2001), 1107–1113.
- [24] Penelope Eckert. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12, 4 (2008), 453–476.
- [25] Penelope Eckert and Sally McConnell-Ginet. 2003. *Language and gender*. Cambridge University Press.
- [26] European Commission. [n. d.]. NUTS—Nomenclature of Territorial Units for Statistics. <https://ec.europa.eu/eurostat/web/nuts/background>.
- [27] Manaal Faruqi, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of EMNLP*.
- [28] Daniel Gayo-Avello. 2012. “I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper”—A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441* (2012).
- [29] Daniel Gayo-Avello. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review* 31, 6 (2013), 649–679.
- [30] Nicholas Genourous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle, and Reid Priedhorsky. 2014. Global Disease Monitoring and Forecasting with Wikipedia. *PLOS Computational Biology* 10, 11 (11 2014), 1–16. <https://doi.org/10.1371/journal.pcbi.1003892>
- [31] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2008. Detecting influenza epidemics using search engine query data. *Nature* 457 (nov 2008), 1012. <https://doi.org/10.1038/nature07634>
- [32] Rob Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching Text: Abstract Features for Cross-lingual Gender Prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 383–389.
- [33] Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers’ age and gender. In *Proceedings of ICWSM*.
- [34] Scott A. Hale. 2014. Global Connectivity and Multilinguals in the Twitter Network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’14)*. ACM, New York, NY, USA, 833–842. <https://doi.org/10.1145/2556288.2557203>
- [35] Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *Proceedings of ICWSM*. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114>
- [36] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [37] D. Holt and T. M. F. Smith. 1979. Post Stratification. *Journal of the Royal Statistical Society. Series A (General)* 142, 1 (1979), 33–46. <http://www.jstor.org/stable/2344652>
- [38] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely Connected Convolutional Networks. In *CVPR*, Vol. 1. 3.
- [39] Aaron Jaech and Mari Ostendorf. 2015. What your username says about you. *arXiv preprint arXiv:1507.02045* (2015).
- [40] Bernard J Jansen. 2010. *Use of the internet in higher-income households*. Pew Research Center Washington, DC.
- [41] Isaac Johnson, Connor McMahon, Johannes Schöning, and Brent Hecht. 2017. The Effect of Population and Structural Biases on Social Media-based Algorithms: A Case Study in Geolocation Inference Across the Urban-Rural Spectrum. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 1167–1178.
- [42] Soon-Gyo Jung, Jisun An, Haewoon Kwak, Joni Salminen, and Bernard J Jansen. 2017. Inferring Social Media Users’ Demographics from Profile Pictures: A Face++ Analysis on Twitter Users. In *Proceedings of The 17th International Conference on Electronic Business*. 140–145.
- [43] Andreas Jungherr. 2017. Normalizing digital trace data. *Digital Discussions: How Big Data Informs Political Communication* (2017).
- [44] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. 2012. Why the pirate party won the german election of 2009 or the trouble with predictions: Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. “Predicting elections with twitter:

- What 140 characters reveal about political sentiment”. *Social science computer review* 30, 2 (2012), 229–234.
- [45] Shari Kendall, Deborah Tannen, et al. 1997. Gender and language in the workplace. *Gender and Discourse*. London: Sage (1997), 81–105.
- [46] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-Aware Neural Language Models. In *AAAI*. 2741–2749.
- [47] Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 108–113.
- [48] Klaus Krippendorff. 2011. Computing Krippendorff’s alpha-reliability. *University of Pennsylvania Departmental papers (ASC)* (2011).
- [49] Robin Tolmach Lakoff and Mary Bucholtz. 2004. *Language and woman’s place: Text and commentaries*. Vol. 3. Oxford University Press, USA.
- [50] Fabio Lamanna, Maxime Lenormand, Maria Henar Salas-Olmedo, Gustavo Romanillos, Bruno Gonçalves, and José J Ramasco. 2018. Immigrant community integration in world cities. *PLoS one* 13, 3 (2018), e0191612.
- [51] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science* 343, 6167 (2014), 1203–1205.
- [52] Roderick Little and Donald Rubin. 2002. *Statistical analysis with missing data, Second edition*. 408 pages.
- [53] Michael McCandless. 2010. Accuracy and performance of Google’s compact language detector. <http://blog.mikemccandless.com/2011/10/accuracy-and-performance-of-googles.html>.
- [54] James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *Proceedings of ICWSM*. 650–653.
- [55] Márton Mestyán, Taha Yasseri, and János Kertész. 2013. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *PLOS ONE* 8, 8 (08 2013), 1–8. <https://doi.org/10.1371/journal.pone.0071226>
- [56] Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. *Proceedings of ICWSM* 11, 5th (2011), 25.
- [57] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.
- [58] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. “How Old Do You Think I Am?” A Study of Language and Age in Twitter. In *Proceedings of ICWSM*.
- [59] Dong Nguyen, Noah A Smith, and Carolyn P Rosé. 2011. Author age prediction from text using linear regression. In *Proc. of the Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, 115–123.
- [60] David K Park, Andrew Gelman, and Joseph Bafumi. 2004. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis* 12, 4 (2004), 375–385.
- [61] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.
- [62] Francisco Rangel, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. 2018. Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter. *Working Notes Papers of the CLEF* (2018).
- [63] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proc. of the 2nd International Workshop on Search and Mining User-generated Contents*. ACM, 37–44.
- [64] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. (2018).
- [65] Sara Rosenthal and Kathleen McKeown. 2011. Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In *Proc. ACL*. Association for Computational Linguistics, 763–772.
- [66] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)* (July 2016).
- [67] Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346, 6213 (2014), 1063–1064.
- [68] Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H. Andrew Schwartz. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proc. EMNLP*. Association for Computational Linguistics, 1146–1151.
- [69] Carl-Erik Särndal and Sixten Lundström. 2005. *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Ltd, Chichester, UK. 1–199 pages. <https://doi.org/10.1002/0470011351>
- [70] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 6. 199–205.
- [71] Luke Sloan. 2017. Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media + Society* 3, 1 (2017), 2056305117698981.
- [72] Deborah Tannen. 1991. *You just don’t understand: Women and men in conversation*. Virago London.
- [73] Deborah Tannen. 1993. *Gender and conversational interaction*. Oxford University Press.
- [74] UK Office for National Statistics 2015. *Harmonised Concepts and Questions for Social Data Sources - Primary Principles*. UK Office for National Statistics. <http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/demographic-information--household-composition-and-relationships.pdf>
- [75] United States Department of Education 2009. *Implementation guidelines: Measures and methods for the national reporting system for adult education*. United States Department of Education. http://www.air.org/sites/default/files/downloads/report/ImplementationGuidelines_0.pdf
- [76] Nicola Van Rijbergen, Katarzyna Jaworska, Guillaume A Rousselet, and Philippe G Schyns. 2014. With age comes representational wisdom in social signals. *Current Biology* 24, 23 (2014), 2792–2796.
- [77] Wei Wang, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. Forecasting elections with non-representative polls. *International Journal of Forecasting* 31, 3 (2015), 980–991.
- [78] Zijian Wang and David Jurgens. 2018. It’s going to be okay: Measuring Access to Support in Online Communities. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 33–45.
- [79] Zach Wood-Doughty, Praateek Mahajan, and Mark Dredze. 2018. Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*. 56–61.
- [80] Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation* 21, 3 (2007), 165–181.
- [81] Emilio Zagheni, Ingmar Weber, and Krishna Gummadi. 2017. Leveraging Facebook’s Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review* 43, 4 (2017), 721–734.
- [82] Jinxue Zhang, Xia Hu, Yanchao Zhang, and Huan Liu. 2016. Your Age Is No Secret: Inferring Microbloggers’ Ages via Content and Interaction Analysis. In *Proceedings of ICWSM*. 476–485.