Correcting Sociodemographic Selection Biases for Population Prediction from Social Media

Salvatore Giorgi,[1] Veronica E. Lynn,[2] Keshav Gupta,[2] Farhan Ahmed,[2]
Sandra Matz,[3] Lyle H. Ungar,[1] H. Andrew Schwartz[2]
[1] University of Pennsylvania
[2] Stony Brook University
[3] Columbia University
sgiorgi@sas.upenn.edu, has@cs.stonybrook.edu

Abstract

Social media is increasingly used for large-scale population predictions, such as estimating community health statistics. However, social media users are not typically a representative sample of the intended population — a "selection bias". Within the social sciences, such a bias is typically addressed with *restratification* techniques, where observations are reweighted according to how under- or over-sampled their socio-demographic groups are. Yet, restratifaction is rarely evaluated for improving prediction.

In this two-part study, we first evaluate standard, "out-of-the-box" restratification techniques, finding they provide no improvement and often even degraded prediction accuracies across four tasks of esimating U.S. county population health statistics from Twitter. The core reasons for degraded performance seem to be tied to their reliance on either sparse or shrunken estimates of each population's socio-demographics. In the second part of our study, we develop and evaluate ROBUST POSTSTRATIFICATION, which consists of three methods to address these problems: (1) *estimator redistribution* to account for shrinking, as well as (2) *adaptive binning* and (3) *informed smoothing* to handle sparse socio-demographic estimates. We show that each of these methods leads to significant improvement in prediction accuracies over the standard restratification approaches. Taken together, ROBUST POSTSTRATIFICATION enables state-of-the-art prediction accuracies, yielding a 53.0% increase in variance explained (R^2) in the case of surveyed life satisfaction, and a 17.8% average increase across all tasks.

Introduction

Digital language has shown promise for inexpensive large-scale population measurement (Coppersmith et al. 2015; Mowery et al. 2016). Twitter, for example, has been used to track public opinion (O'Connor et al. 2010; Miranda Filho, Almeida, and Pappa 2015) and measure community health (Mowery et al. 2016; Abebe et al. 2020; De Choudhury, Counts, and Horvitz 2013). The passive assessment of community characteristics that are otherwise expensive to obtain offers tremendous opportunities for both researchers and practitioners, but it also poses a challenge that is often overlooked: predictions made from social media are often prone to significant bias resulting from non-representative samples.

Although the user bases of social media platforms are diversifying, they do not accurately reflect the general population (Duggan and Smith 2013; Greenwood, Perrin, and Duggan 2016). For example, Twitter users typically are younger and have a higher median income (Duggan and Smith 2013). Per location, such as counties in the U.S., these biases can further differ. As a result, samples collected from Twitter are not representative of the populations they are intended to model, leading to a "selection bias" that can potentially skew results.

In this study, we address the issue of selection bias when using spatially aggregated Twitter language to measure community health and well-being. We estimate age, gender, income, and education distributions of a geolocated Twitter sample through pretrained socio-demographic models. When compared to known distributions of the community (via the U.S. Census), these inferred socio-demographic variables allow us to quantify the selection bias per observation (a county in this case). Using these insights, we estimate county-level language features, weighting each county member according to how over- or under-represented they are within the community's known socio-demographic distribution.

While addressing selection bias is a common procedure in many quantitative social sciences, it is primarily used to improve in-sample correlational statistics (Berk and Ray 1982; Winship and Mare 1992). In contrast, attempts to address selection bias to improve predictive models (i.e. supervised NLP) are rare. One potential explanation for this gap is that socio-demographic information is rarely available in predictive contexts, which rely on observable data rather then self-report questionnaires. However, recent work demonstrates that estimates of demographics from language can be highly associated with self-reported demographics (Zhang et al. 2016; Chen et al. 2015), which could provide a means to estimating and correcting demographic selection bias.

Similar to in-sample corrections of selection biases, one would expect that the accuracy of predictive models (e.g., predictions of **representative** health outcomes) can be improved by taking account of observable selection biases. Here, we show that this is not the case: applying standard in-sample solutions (e.g., post-stratification and raking) leads to a decrease in performance when predicting representative county health. Upon investigation, we identify that this drop in performance arises from two problems: (1) the use of estimated socio-demographics and (2) the sparse socio-demographics bins (i.e., socio-demographic subgroups which are sparsely populated in a sample when compared to known populations). Building on these findings, we propose novel solutions to each of these problems in the form of (1) estimator redistribution and (2) adaptive binning and informed smoothing.

This paper presents methods and results in two stages. First, we define and contextualize the problem of selection bias, presenting existing methods (**Out-of-the-box Correction Techniques**) and highlighting the fact that standard methods fall short in our setup (**Results based on Existing Methods**). Second, we introduce novel methods to handle challenges common to selection bias correction in **Improving Correction Techniques** and present results in **Results: Estimation and Sparsity Challenges**.

Contributions Our key contributions include: (1) Introduce the problem of selection bias correction for supervised NLP, including standard methods from other fields; (2) Show that standard reweighting techniques used widely in other fields often lead to degraded performance in predicting health statistics from social media language; (3) Identify the problems of standard reweighting and develop methods to mitigate them; (4) Apply these techniques as ROBUST POSTSTRATIFICATION to obtain state-of-the-art prediction accuracies of county life satisfaction and health. We also open-source all code.[1] and data[2]

Problem Statement: Selection Bias Correction[3]

Given hierarchical data where lower level individual data points (i.e., Twitter users) are nested within a population (i.e., U.S. county), we wish to estimate the representative population-level expectation, μ_{X_i}, from the lower level data. For example, when correcting for selection bias of language on Twitter, X is a vector of linguistic features for which we wish to derive a representative mean.

Simple averaging methods fail to account for differences between the observed sample (individuals in our Twitter data mapped to a county) and the target population (the entire population of a county) for whom a measurement is desired. Thus, with respect to a target population, the measurements over the sample are biased, i.e. suffer a selection bias. More formally, we define $d = \{d_m\}$ to be a set of individual level auxiliary variables (in our case, $d = \{$age, gender, income, education$\}$), $\mathrm{Q}_i(d)$ to be the distribution of our sample (those for whom we have a measurement in our data set) and $\mathrm{P}_i(d)$ to be the distribution of the target population (those for whom a measurement is desired) in U.S. county i. Then, following Shah, Schwartz, and Hovy (2020), we take selection bias to mean that the sample distribution is dissimilar from a theoretically-desired distribution (the census-measured population distribution in this case):

$$\mathrm{P}_i(d) \nsim \mathrm{Q}_i(d).$$

	Definition
$d^{(s)}$	Twitter user's predicted socio-demographics value in the Twitter sample s
$d^{(t)}$	Twitter user's redistributed socio-demographics value in the target distribution t
d_m	Twitter user's predicted value for socio-demographic m (age, gender, income, education)
D_{d_m}	Partition of the socio-demographic d_m
$D_{d_m}^{(h)}$	Subset of partition D_{d_m}
i	Index, population level observations (U.S. counties)
j	Index, individual level observations (Twitter users)
m	Index, socio-demographics (age, gender, income, education)
h	Index, socio-demographic partitions
d	Set of socio-demographic variables (subsets of $\{$age, gender, income, education$\}$)
k	Smoothing parameter
$\min_h^{(s)}$	Minimum socio-demographic value in subset $D_{d_m}^{(h)}$ in our Twitter sample s
$\max_h^{(s)}$	Maximum socio-demographic value in subset $D_{d_m}^{(h)}$ in our Twitter sample s
$\min_h^{(t)}$	Minimum socio-demographic value in subset $D_{d_m}^{(h)}$ in our target distribution t
$\max_h^{(t)}$	Maximum socio-demographic value in subset $D_{d_m}^{(h)}$ in our target distribution t
N_i	Cardinality of U_i (number of Twitter users in county i)
$\mathrm{P}_i(d)$	Distribution of the target population (U.S. Census)
$\mathrm{Q}_i(d)$	Distribution of our sample (Twitter sample)
s	Sample distribution (Twitter users)
t	Target distribution (U.S. Census)
U_i	Set of all individuals within a population (Twitter users county i)
X_i	Population level observation (county level linguistic feature)
$x_{i,j}$	Individual level observation (Twitter user linguistic feature)
$\psi_i(d)$	Correction factor
μ_{X_i}	County level expectation of X_i

Table 1: Definitions for notation used throughout the paper.

We can then view selection bias correction as estimating a correction factor for the given set of auxiliary variables d:

$$\psi_i(d) = \frac{\mathrm{P}_i(d)}{\mathrm{Q}_i(d)}, \tag{1}$$

such that our goal of estimating μ_{X_i}, the population expectation of the individuals' features X_i for community i, can be written as

$$\hat{\mu}_{X_i} = \frac{1}{N_i} \sum_{j \in U_i} \psi_i(d) r_j(x_j). \tag{2}$$

Here U_i is the set of individuals in community i, with $N_i = \left|U_i\right|$, and r_j is some kernel function. Note that Eq. 1 is similar to the Kullback-Leibler divergence (Kullback and Leibler 1951). Since we would like a multiplicative correction factor, we do not take the log of the ratio.

This formulation, rooted in the literature on reweighting and post-stratification techniques from economics and social science (Kalton and Flores-Cervantes 2003; Hoover and Deghani 2020), includes several useful abstractions. First, d

[1] Code: https://github.com/wwbp/robust-poststratification

[2] Data and Supplemental Materials: https://osf.io/ae5w6

[3] See Table 1 for definitions of all terms.

can contain any number of auxiliary variables, though, here d is a set of socio-demographics of the Twitter users (e.g., various combinations of age, gender, income, and education). Higher level populations need not be limited to counties (e.g., cities and countries), nor do these populations need not be limited to spatial regions. Further, although our focus is social media-based community measurements, this formulation could be used for other types of data: estimating consumer metrics per household (Alexander 1987), state polling (Park, Gelman, and Bafumi 2004), and, generally, individual-level data from a biased sample of a population.

Bias in Social Media Population Measurement

Samples collected from Twitter aren't generally representative of the real-world populations that they are intended to model (Mislove et al. 2011; Culotta 2014). To some extent, this is attributable to unbalanced user demographics – users skew young, toward one gender or the other, toward wealthy or poor, and toward urban rather than rural (Duggan and Smith 2013; Greenwood, Perrin, and Duggan 2016; Hecht and Stephens 2014). Beyond demographics of who "selects" to use social media, data collection methods further contribute to selection biases. The geotagging process can select certain ages and genders (Pavalanathan and Eisenstein 2015), and races may be partially excluded due to language filters, prone to errors on region- or race-specific dialects such as African-American English (Blodgett, Green, and O'Connor 2016).

Limited work has been done to correct for selection biases on social media. Recently, Wang et al. (2019) presented a method for selection bias correction to create national population estimates from social media. They showed that one could use inferred demographics with a traditional post-stratification technique to produce more representative population statistics. We also use estimated demographics, but we find these traditional post-stratification techniques have problems which lead to degraded performance for predictive modeling. Other fields have presented social media-specific frameworks (Zagheni and Weber 2015) but without predictive evaluations.

While population studies often attempt to correct for selection bias, few have explored the use of corrections to improve predictive modeling. Non-representative samples can have a significant impact on model performance. For example, Weeg et al. (2015) used mentions of diseases on Twitter and nearly doubled predicting prevalence rates for 22 diseases after limiting analysis to disease prevalence amongst known Twitter users. Using Twitter to predict elections, Miranda Filho, Almeida, and Pappa (2015) explored selection bias as a reason for inconsistent election predictions. Attempting to construct stratified samples, they concluded that results were encouraging but lacked sufficient data to make predictions. Our method works even in cases such as this where traditional restratification isn't feasible.

Closest to our work, we build on ideas from Culotta (2014) who explored reweighting schemes for predicting county-level health statistics. They reweighted instances according to users' predicted gender and race, leading to improved predictions for 20 out of 27 variables. However, their evaluation was limited to the top 100 most populous counties, which are primarily homogeneous urban centers. In contrast, our work explores methods for cases where the data is not homogeneous and/or when data is sparse. Further, we provide a more comprehensive evaluation and correct for more variables (e.g. age, gender, education, and income).

Data[2]

Our training data is broken into two pieces: (1) the biased sample from tweets and (2) representative population-level survey percentages. The biased sample consists of Twitter data which we want to aggregate to the community level in such a way that its socio-demographic makeup matches that of our representative population, the U.S. Census.

Biased Sample: Twitter We use the open source County Tweet Lexical Bank — a U.S. county-mapped Twitter data set built over 1.6 billion tweets (Giorgi et al. 2018). The County Tweet Lexical Bank Twitter data was pulled from July 2009 to February 2015, geolocated to U.S. counties (i.e., county FIPS codes, which are unique numeric identifiers for U.S. counties) via self-reported location information in public account profiles and latitude / longitude coordinates (see Schwartz et al. (2013a) for the county mapping/geolocation process) and then filtered to contain only English tweets (Lui and Baldwin 2012). At a high level, the county mapping process is as follows: (1) if a tweet object contains latitude / longitude coordinates, we can trivially map that tweet to a county, (2) if a Twitter account has self-reported location information in the profile (e.g., "'living in NYC") we match the location string in the profile to U.S. cities which can be mapped to counties. The data was then limited to Twitter accounts with at least 30 posts and U.S. counties represented by at least 100 such unique accounts. The final Twitter data set consists of 2,041 U.S. counties with 6.06 million users.

Representative Population: U.S. Census Five year estimates (2011-2015) for age, gender, education, and income were obtained from United States Census Bureau's 2015 American Community Survey (ACS). Age records contain the percentages of people within age ranges 18-19, 20-24, 25-29, 30-34, 35-39, 40-44, 46-49, 50-54, 55-59, 60-64, and above 65. The gender records consist of the percentages of males and females for each county. Percentages of income for the following bins: less than $10,000, $10,000-$14,999, $15,000-$24,999, $25,000-$34,999, $35,000-$49,999, $50,000-$74,999, $75,000-$99,999, $100,000-$149,999, $150,000-$199,999, and greater than $200,000. Education is divided into two groups: percentage of the population with less than a Bachelor's degree and percentage higher than that of Bachelors.

Outcomes Selection bias correction is evaluated across four different community level out-of-sample prediction tasks, where we cross sectionally predict county level health variables (i.e., the outcomes or dependent variable in each task) from county level language on Twitter (i.e., the independent variables). The four tasks include two measures of objective health (heart disease and suicide mortality rates) and two measures of subjective health and well-being (Life

Satisfaction and Poor or Fair Health). From the Centers for Disease Control and Prevention (CDC) we collected age-adjusted mortality rates for heart disease ($N = 2,038$) and suicide ($N = 1,672$), averaged across 2010-2015. Life satisfaction scores are calculated as average individual level response to the question "In general, how satisfied are you with in your life?" (1 = very dissatisfied and 5 = very satisfied), averaged across 2009 and 2010 ($N = 1,951$) (Lawless and Lucas 2011). Finally, Poor or Fair Health was obtained from the County Health Rankings and is sourced from the Behavioral Risk Factor Surveillance System (BRFSS) (Remington, Catlin, and Gennuso 2015). This is an age-adjusted measure of the percentage of adults who consider themselves to be in poor or fair health (i.e., percentage of adults who answered fair or poor to the question: "In general, would you say that in general your health is Excellent/Very good/Good/Fair/Poor?"; $N = 1,931$). All N reported above are a subset of the 2,041 counties with sufficient Twitter data, and are therefore the final number of observations in our four county tasks.

PEW: National Social Media and Twitter Use Since Twitter socio-demographic populations are not known at the county level we use National statistics collected from PEW's Social Media update (Greenwood, Perrin, and Duggan 2016), including age, gender, income, and education for the years 2013-2016. Demographics were binned as follows: age (18-29, 30-49, 50-64, and 65+), gender (female / male), income (less than $30,000, $30,000-$49,999, $50,000-$74,999, and $75,000+; yearly), and education (high school grad or less, some college and college+). For each demographic bin we collect the percentage of the population who use social media and the percentage of the population who use Twitter. These percentages are averaged over the four years available, the closest available timeline to the Twitter sample. Using the total U.S. population we then calculate the percentage of people in each socio-demographic bin (i.e., total U.S. population × average bin percentage of people who use social media × average bin percentage of people on Twitter).

Ethics Statement This study was reviewed and approved by an academic institutional review board, found to be exempt, non-human subjects data, with *none to minimal* chance of harm to individuals. All raw data used in this study are publicly available. Our aggregate anonymized (manually checked for identifying information) language features by county are publicly available.[2] For additional privacy protection, **no** individual-level estimates, intermediate information derived within the approach, will be made available. The original tweets, which are publicly available, are not able to be redistributed due to Twitter's Terms of Service.

Estimating Socio-demographic Bias from Language

Sample socio-demographics are necessary in order to quantify and correct non-representation, but such information is not typically available in social media. We thus turn to socio-demographic estimates of our sample from their language. Such estimates have been validated in a number of contexts including Twitter (Sap et al. 2014; Matz et al. 2019),[4] and a similar approach was used by Wang et al. (2019). We produced language-based estimates for four socio-demographic variables, which we will correct for selection bias: age, gender, income, and education. All four estimators are described below. The median (or percentage) county values for our sample estimates versus census population statistics are given in Table 2. On average, our Twitter sample appears younger and more educated than the population as a whole, but it is important to remember bias may differ from one county to the next and correction attempts to make each county more representative of its population. We note that gender is fairly evenly split across U.S. counties across all three categories: U.S. Census, Twitter sample, and PEW.

	Age	Perc. Female	Income	Perc. Bachelor's Degree
Census	39.3	50.4	$48,280	22.3
Twitter	22.1	53.8	$36,437	40.5
PEW	28.8*	48.3	$58,660*	42.9

Table 2: County mean of medians or percentages. *Imputed from bin percentages as median not provided by Pew.

We utilized estimated age, gender, income, and education, based on tweet language, using the following models.[5]

Age and Gender. Age and gender estimates were based on a demographic predictive lexica (Sap et al. 2014). Sap inferred these models over a set of annotated users with self-reported age and gender (binary as multi-class gender was not available at the time) from Facebook, Twitter, and blogs. Accuracy of the estimates for age correlated with self-reported age at Pearson r = 0.86 and the estimated gender with an accuracy = 0.90 with self-reported gender. The model produced real values for age which were thresholded to between 13 and 80. For gender, output was a continuous score from negative (more male) to positive (more female). Because county statistics were limited to binary gender these were converted to 1 for "female" and 0 otherwise.

Income. Income was estimated using the model built in Matz et al. (2019). They collected a sample of 2,623 participants from Qualtrics in 2015 who reported their annual income and shared social media language. This model achieved an out-of-sample accuracy of Pearson r = .41 for estimated income as compared to true income. The model takes ngram frequencies as well as social media topic loadings from Schwartz et al. (2013b) as input.

Education. An education classification model was built over a sample of users recruited from Qualtrics (Preoţiuc-Pietro et al. 2017). A total of 4,062 users reported education level and shared their Facebook status data. Mirroring Matz's income model, for each user, we extracted ngrams of length 1 to 3 and loadings for a set of 2,000 social media-based LDA topics (Schwartz et al. 2013b). We used a multi-class

[4] While perfection is not necessary to achieve benefit, excessive error would presumably prevent our approach from improving county-level predictions.

[5] Age, gender, and income estimation models were previously published, while education is novel to this paper.

linear-svc classifier and train on the following classes: (1) less than high school diplomam (2) high school diploma or Associate's degree, and (3) Bachelor's degree or higher. This model obtained an accuracy of .62 and an F1 score of .53 using 10-fold cross validation. We then used this model to predict class probabilities for Twitter accounts and, because most county data only indicated higher education percentages, collapsed the first two classes into a single class. This resulted in two final education classes: (1) less than a Bachelor's degree and (2) Bachelor's degree or higher.

Study 1: Out-of-the-box Correction Techniques

Our approach to applying standard selection bias correction relies on three steps: (1) estimating socio-demographics, (2) creating weight factors, and (3) reweighting user level features and aggregating to the county (i.e., applying weight factors).

In practice $\mathrm{P}_i(d)$ and $\mathrm{Q}_i(d)$ are unknown and must be estimated, typically by creating a partition D_{d_m} of each socio-demographic variable d_m into non-overlapping subsets $D_{d_m}^{(h)}$ where $\bigcup_h D_{d_m}^{(h)} = D_{d_m}$:

$$\hat{\psi}_i(d) = \frac{\mathrm{P}_i(d|d_m \in D_{d_m}^{(h)}, \forall m)}{\mathrm{Q}_i(d|d_m \in D_{d_m}^{(h)}, \forall m)}. \tag{3}$$

Furthermore, the population distribution $P_i(d)$ is estimated using population percentages from known national surveys, in our case, the U.S. Census, and the sample distribution $Q_i(d)$ is estimated from our sample percentages:

$$\hat{\psi}_i(d) = \frac{\mathrm{perc}_{\mathrm{pop}}(d|d_m \in D_{d_m}^{(h)}, \forall m)}{\mathrm{perc}_{\mathrm{samp}}(d|d_m \in D_{d_m}^{(h)}, \forall m)}, \tag{4}$$

where $\mathrm{perc}_{\mathrm{pop}}$ and $\mathrm{perc}_{\mathrm{samp}}$ are the population and sample percentages, respectively. The non-overlapping subsets $D_{d_m}^{(h)}$ are referred to as bins throughout.

Existing Methods

We investigate two common methods for creating weight factors: (1) naive post-stratification and (2) raking, both of which are a form of post-stratification. These two methods can be viewed as different ways of estimating the joint probability distribution in the population domain of the given socio-demographic d: $\mathrm{P}_i(d)$ from Equation 1.

Post-stratification. Post-stratification reweights each user according to the joint distribution of a set of socio-demographics (Holt and Smith 1979; Little 1993; Henry and Valliant 2012). In practice, this joint distribution is rarely known or available to researchers beyond two or three variable combinations. The two methods below address this situation and use only the marginal distributions for each socio-demographic.

Naive Post-stratification. Since joint distributions are not always available for many variables of interest, one can estimate the joint distribution from given marginals. One approach is to assume all marginal distributions are independent (Leemann and Wasserfallen 2017). This method multiplies the proportion of people in each marginal bin to estimate the proportion of people in each of the joint distribution's bins, mirroring the assumption of Naive Bayes ($p(a, b) = p(a)p(b)$).

Raking. Raking is an iterative method which operates on the marginal distributions, adjusting each sample marginal to match the population distributions (Deville, Särndal, and Sautory 1993). For example, raking over age and gender would first adjust age sample marginals to match age population marginals, and then adjust gender sample marginals to match gender population marginals. This process is repeated until the marginal distributions of the sample variables match the population marginal distributions within some small margin of error. The adjusted sample marginals are then substituted into the numerator of Equation 1.

Applying Weight Factors and Predictive Modeling

We apply our correction weights to individual level (Twitter users) linguistic features, specifically the top 25,000 most frequent unigrams across our entire sample, noting that this procedure will work for any individual level data. We concatenate all tweets from each user in our data set and tokenize using a tokenzier built for social media data (Schwartz et al. 2017). We then encode each unigram the relative frequency of use for each given user. Using Equation 2, each linguistic feature x_j is aggregated from user j to county i:

$$\hat{\mu}_{X_i} = \frac{1}{N_i} \sum_{j \in U_i} \hat{\psi}_i(d) r_j(x_j). \tag{5}$$

Here U_i is the set of users in county i, N_i is the total number of Twitter users in county i, $\hat{\psi}_i(d)$ is the correction weight of the demographic set d, and $r_j(x_j)$ is the relative frequency of the unigram x_j for user j. When aggregating from user to county with no bias correction we set $\hat{\psi}_i(d) = 1, \forall j \in U_i$ and $\forall i$. The end results is a set of 25,000 county-level average unigrams.

Predictive Modeling

Since our methods focus is on selection bias, we integrate our correction approach into an established approach for estimating county-level health statistics from language (Eichstaedt et al. 2015; Giorgi et al. 2018; Jaidka et al. 2020).

Features The county-level average unigrams are then used to derive a set of topic loadings for each county. We use a set of 2,000 topics derived from Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). The topics were built over the myPersonality Facebook data set, which consists of approximately 15 million; see Schwartz et al. (2013b) for more details on the topic modeling process. These topics have been used across a number of studies predicting county-level health and well-being (Eichstaedt et al. 2015; Giorgi et al. 2018; Curtis et al. 2018).

Modeling For each of our four county-level prediction tasks, we predict the outcome (i.e., heart disease mortality, suicide mortality, life satisfaction, and percentage in poor or fair health) using the 2,000 topic features described above in

	Heart Disease	Suicide	Life Sat.	Poor/Fair Health	Avg.
Baseline	.751	.614	.445	.748	.640
Age	$.622^-$	$.492^-$	$.239^-$	$.595^-$	$.487^-$
Gender	$.755^+$	$.619^+$	$.437^-$	.744	.641
Income	$.703^-$	$.530^-$	.455	$.716^-$	$.599^-$
Education	$.762^+$	.617	$.457^+$	$.754^+$	$.648^+$

Table 3: Standard methods using single correction factors. Reported Pearson r., $^+$ and $^-$ indicate a significant increase and decrease in performance, respectively ($p < 0.05$). Half of the "out-of-the-box" correction factors reduce accuracy.

a 10-fold cross validation setup. Thus, the final data size is dependent on the county outcome (i.e., dependent variable), with 2,000 LDA topics (i.e., independent variables) constant across each task: heart disease mortality $N = 2,038$; suicide mortality $N = 1,672$; life satisfaction $N = 1,951$; and percentage in poor or fair health $N = 1,931$. The topic features are then fed through a three step feature selection pipeline. First, we remove all low variance features. Next, we remove features which are not correlated with our county-level outcomes at a family-wise error rate α of 60. At this point, all features are standardized: mean centered and normalized by the standard deviation. Finally, stochastic principal component analysis is applied to the topic features in order to reduce the size of the feature set to approximately 10% its original size. For each of the 10 folds, we train an ℓ_2 penalized ridge regression (Hoerl and Kennard 1970) on 9 of the folds and apply the model to the held out 10th fold. All models use a regularization term λ of 10,000. Similarly to the topic feature set, this same pipeline has been successfully used across a number of county-level health studies (Eichstaedt et al. 2015; Giorgi et al. 2018; Jaidka et al. 2020; Abebe et al. 2020)

Results for Out-of-the-box Methods

In this section we evaluate how well existing post-stratification techniques improve prediction accuracy by correcting for selection bias. We focus on the average cross-validation accuracy across the four health outcomes introduced previously: heart disease mortality, suicide mortality, life satisfaction, and percent in poor or fair health. The assumption is that if a mitigation technique is useful it should improve predictive performance, while unnecessary or erroneous techniques will have no or negative effect.

Predictive Performance As shown in Table 3, we found a decrease in performance (Pearson r) when attempting to correct for both age and income biases. Gender correction has mixed results, as we see an increase for heart disease and suicide, a decrease for life satisfaction and no change for poor/fair health. Education gives a significant boost for three out of four tasks, with no change for suicide. For each correction factor we perform a pair t-test on the model's residual as compared to the baseline model residual. We report both positive ($^+$) and negative ($^-$) statistical differences at $p < 0.05$.

Baseline $= .640$	Naive Post-Stratification	Raking
Age + Gender	$.514^-$	$.473^-$
Income + Education	$.600^-$	$.591^-$
Age + Gen. + Inc. + Edu.	$.627^-$	$.541^-$

Table 4: Standard methods using multiple correction factors, average predictive accuracy (Pearson r) across four tasks, $^-$ significant decrease in performance at $p < 0.05$. All six "out-of-the-box" combinations reduce predictive accuracy.

Similar patterns hold when averaging across all four tasks. We note that at this point in the paper we will report the average Pearson r across all four tasks (heart disease, suicide, life satisfaction, and poor/fair health). Additionally, tests for significance are done by combining the dependent p-values across the four county-level tasks using the methods developed by Kost and McDermott (Kost and McDermott 2002). We also note that age and income have 10 and 11 possible bins, respectively, whereas both gender and education are binary variables. An increased number of bins can lead to more extreme weights if any bins are densely or sparsely populated, thus increasing the noise in our model. This suggests some issues perhaps arising from having many bins (e.g. sparse or unstable estimates of people per bin; we will address this in the next two sections).

Average predictive performance for combinations of correction factors is given in Table 4. Again, across the board we see no increase in predictive performance when comparing to baseline. Additionally, we see no increase in predictive performance when comparing to a single factor correction in Table 3.

Study 2: Robust Poststratification for Improved Correction

Standard selection bias mitigation techniques not only provided no benefit, but, on average, tended to hurt performance within the context of predicting county health from social media language. We hypothesize this is due to two challenges. First, socio-demographic estimation from language introduces systematic effects to the distributions (e.g. from shrinkage – bias toward the mean). Second, data sparsity is an issue when dealing multi-dimensional socio-demographics, since some counties contain as few as 100 individuals.

Methods

Challenge 1: Estimator Shrinking The first challenge originates in Step 1 of our pipeline: estimating socio-demographics from text. The estimators used to create the linguistic socio-demographic scores are regularized which shrinks the estimated distribution towards the mean of the training data. To compensate for this, each users' estimated socio-demographics are redistributed such that our source distribution matches that of a target distribution, in our case, that of the national distribution of Twitter users ($expectedBinPercs$, as reported by the PEW Reseach Center). ***Estimator redistribution*** shifts each acount's linguistic socio-demographic estimates such that the population percentage in each source bin matches those of the target bins.

Specifically, for a given socio-demographics bin h, the bin boundaries in the source data ($\min_h^{(s)}$ and $\max_h^{(s)}$) were determined such that they match proportions in target population distribution bins ($\min_h^{(t)}$ and $\max_h^{(t)}$). See Algorithm 1 for details and the Supplemental Materials[2] for a step-by-step example of this method.

A given user's estimated socio-demographic $d^{(s)}$ (where s is the source distribution) is redistributed using the following equation:

$$\frac{d^{(s)} - \min_h^{(s)}}{\max_h^{(s)} - \min_h^{(s)}} = \frac{d^{(t)} - \min_h^{(t)}}{\max_h^{(t)} - \min_h^{(t)}},$$

The redistributed estimation value is obtained by solving for $d^{(t)}$, the socio-demographic in the target distribution t:

$$d^{(t)} = \left(d^{(s)} - \min_h^{(s)}\right) \frac{\max_h^{(t)} - \min_h^{(t)}}{\max_h^{(s)} - \min_h^{(s)}} + \min_h^{(t)}. \tag{6}$$

Figure 1 shows the age distributions of our national Twitter sample and PEW's reported national percentages.

We expect estimator redistribution to help when there is a large number of socio-demographic bins and when there exists large differences between the sample and target distributions, regardless of the number of bins. The redistribution process will move users from densely populated bins into sparser bins, yielding more stable correction factors — users in extreme bins (either dense or sparse) are severely under or over-weighted.

Challenge 2: Sparse Data Bins The second challenge originates in Step 2 of our pipeline: creating weight factors. As we (1) increase the number of bins of our socio-demographic variables and (2) increase the number of socio-demographic variables we wish to correct for, the probability that any one of our sample users falls into a given bin also shrinks. As seen in Equation 1, as the percentage of users in our sample shrinks, weights will increase. The raking process described above also suffers from the fact that convergence is not guaranteed if empty bins are present (Battaglia et al. 2009). Therefore we focus on ways of estimating $\mathrm{Q}_i(d)$ in Equation 1 such that we mitigate this data sparsity problem.

Algorithm 1: *Estimator Redistribution*

```
Input: $\{d^{(s)}\}$ - demographic estimates from users
       $expectedBinPercs$ - Expected percentages
Output: $\{d^{(t)}\}$ - redistributed demographic estimates
        from users
Def EstRedist ($\{d^{(s)}\}$, $expectedBinPercs$):
    for h in length(expectedBinPercs) do
        p = percentBetween($\min_h^{(t)}$, $\max_h^{(t)}$)
        if l == 0 then
            $\min_h^{(s)}$ = $\min_h^{(t)}$
        else
            $\min_h^{(s)}$ = $\max_{h-1}^{(s)}$
        $\max_h^{(s)}$ = $\min_h^{(s)}$ + 1
        while percentBetween($\min_h^{(s)}$, $\max_h^{(s)}$) $< p$ do
            $\max_h^{(s)}$ += 1
        for $\{d^{(s)}\}$ do
            if $\min_h^{(s)} \le d^{(s)} < \max_h^{(s)}$ then
                $d^{(t)}$ = EquationSix($d^{(s)}$)
    return $\{d_t\}$
```

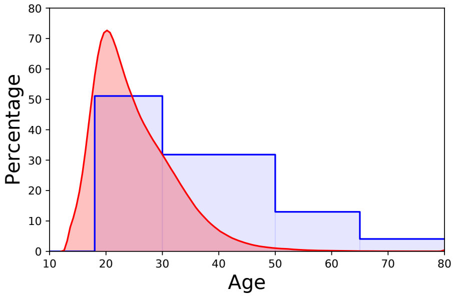


Figure 1: Probability density of the age distributions of our *Twitter sample* (red) versus the expected *Twitter population* distribution according to PEW (blue). Due to regularization which shrinks estimates toward the mean, our age distribution skews significantly younger than PEW.

Adaptive Binning Our first method to account for sparse data sets a minimum threshold on the number of observations within each bin (or partition subset) for a given socio-demographic variable. Adjacent bins are iteratively combined until all bins meet our threshold or we have a single bin. Since both gender and education start with two bins, if either bin fails to meet the threshold then we end up with a single bin and therefore no correction. Thus, we do not expect either variable to significantly increase or decrease predictive performance from baseline. We also note that adaptive binning occurs per socio-demographic (e.g., when correcting for both age and income, we bin age and income separately). While a minimum bin threshold has previously been used (Battaglia et al. (2009) who suggest a minimum bin percentage of 5%), we know of no systematic study of the effect of minimum bin sizes. Additionally, our threshold is set on the number of observations as opposed to a percentage, since percentages will be noisy for sparsely populated counties. The Adaptive Binning algorithm is shown in Algorithm 2; see Supplemental Materials[2] for a step-by-step example of this method.

Informed Smoothing The second method we develop to account for data sparsity uses a smoothing technique that pads each weight with a fraction of users from a known distribution. More formally, we state the source probability in terms of the smoothing constant k as

$$\hat{\mathrm{Q}}_i^{(k)}(d|d_m \in D_{d_m}^{(h)}, \forall m) = \frac{N_s + k\hat{\mathrm{P}}_i(d)}{N_i + k}. \tag{7}$$

Here N_s is the number of sample users with socio-demographic d, N_i is the number of Twitter users in county i

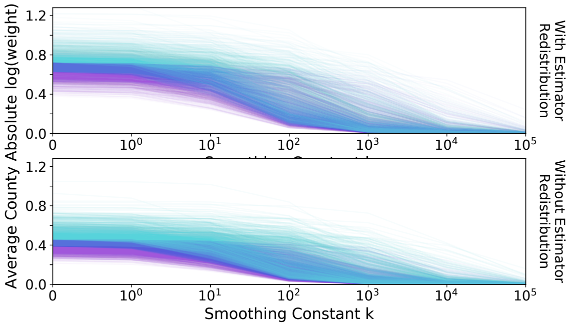


Figure 2: Average, absolute log of the county income weights at different smoothing levels; colored by terciles with (top) and without (bottom) estimator redistribution.

and h is summed over the socio-demographic partition. Note that as $k \to \infty$ as have $\hat{\mathrm{Q}}_i^{(k)}(d) \to \hat{\mathrm{P}}_i(d)$ and therefore all correction weight factors equal 1.

Unlike adaptive binning, informed smoothing does not depend on the total number of bins. Thus, we expect it to have an effect on gender and education correction. This approach is inspired by similar approaches to modeling ngram probabilities in language modelling (Kneser and Ney 1995).

Results: Estimation and Sparsity Challenges

To be sure our methods work as expected, we first observe their effects on the correction weights. Figure 2 shows results for both estimator redistribution and informed smoothing for income alone. Each figure shows the county average, absolute log of the users' correction factors. First, ignoring the effects of smoothing (i.e., focusing on $k = 0$), we see that estimator redistribution shrinks the variance in the correction factors. This is to be expected since estimator redistribution spreads out the distribution — the sample distribution (red) in Figure 1 is spread to match the true distribution (blue). This causes our bins to (1) be less sparse near the tail of the distribution (thus, shrinking large correction factors towards the mean); and (2) less dense near the peak of the distribution (similarly, increasing small correction factors towards the mean).

Algorithm 2: Adaptive Binning

```
Input: $binCounts$ - list of bin counts
       $minBinNumber$ - min integer bin threshold
       $binRanges$ - list of bin ranges
Output: $binCounts$ - list of combined bin counts
        $binRanges$ - list of combined bin ranges
Def AdaptBin ($binCounts, minBinNumber$):
    while min(binCounts) $<$ minBinNumber do
        m = min(binCounts)
        if $m \geq$ minBinNumber then
            break
        i = binCounts.index_of(m)
        combineAdjacent(binCounts[i],
          min(binCounts[i-1],binCounts[i+1]))
        combineAdjacent($binRanges$[i],
          min($binRanges$[i-1],$binRanges$[i+1]))
    return $binCounts$, $binRanges$
```

Baseline $= .640$	Post-Stratification	Naive Post-Stratification	Raking
Age	$.587^+$	-	-
Gender	.639	-	-
Income	$.625^+$	-	-
Education	.648	-	-
Age + Gender	-	$.582^+$	$.584^+$
Inc. + Edu.	-	$.629^+$	$.626^+$
All	-	$.638^+$	$.588^+$

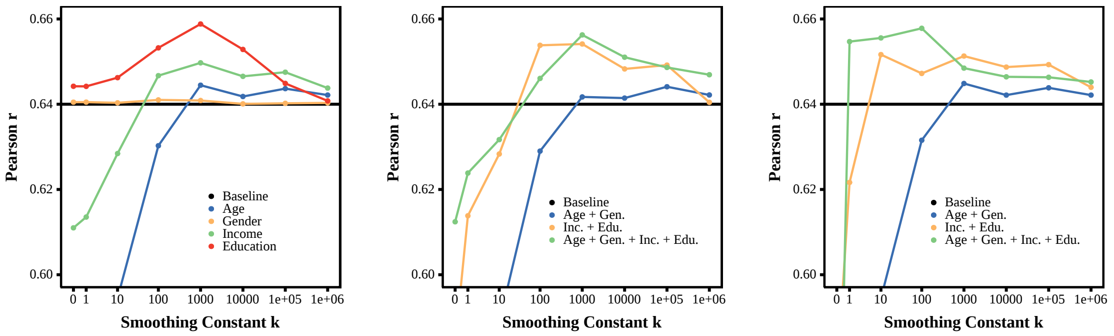
Table 5: Average predictive accuracies (Pearson r) across four tasks when using *estimator redistribution*. $^+$ and $^-$ indicate a significant increase or decrease, respectively, as compared to same correction variable / method pair in Tables 3 and 4. All correction factors, except *gender* and *education*, show significant increases over the "out-of-the-box" methods. "All" includes age, gender, income, and education.

Informed smoothing also has a similar shrinking effect at large k, with average log weights converging to zero. This is expected since Equation 7 says that, as k increases, the estimated sample distribution $\hat{\mathrm{Q}}_i^{(k)}(d)$ matches the estimated population distribution $\hat{\mathrm{P}}_i(d)$. Thus, $\hat{\psi}_i(d) \to 1$ and the log approaches 0. Finally, we see that the variance in weights does not monotonically decrease, with maximum variance at $k = 100$. At $k = 100$ we see the terciles calculated at $k = 0$ spreading out. Since we are correcting weights on a county-by-county basis, with each county having it's own selection bias, we would hope that the average county weights show variance. At $k = 0$ we see this is not the case and all counties have similar average weights, implying that all counties are experiencing the same selection bias (i.e., a similar ratio of population to sample; Equation 1).

Table 5 evaluates the benefit of applying estimator redistribution. Comparing to Tables 3 and 4, we see a marked improvement above post-stratification without estimator redistribution in almost all situations. It does not put us above baseline ($r = .640$) but it is moving in the right direction, so we use estimator redistribution in all remaining experiments.

The predictive accuracies for the adaptive binning experiments are shown in Table 6. This marked our first improvement over the baseline average Pearson r of .640. In most cases, we see a decrease in performance when setting the minimum bin threshold to 1 when compared to no binning. We also see naive combining outperforming raking when k is low, though $k \geq 50$ reverses this and raking outperforms naive. Increasing the minimum bin threshold gradually improved results, with peaks around 100 where all approaches did better than no adaptive binning. As expected, most factors approach baseline when the bin threshold is 1,000.

Figure 3 shows the predictive accuracies of the informed smoothing method. Figure 3(a) shows informed smoothing with single correction factors. For single factors alone, we see



(a) Single Factors (b) Naive Factors (c) Raking Factors

Figure 3: Prediction accuracies when using *informed smoothing*, averaged over all four tasks: Graphs zoomed in to highlight difference near baseline; smoothing constant $k = 0$ is equivalent to no smoothing (see Table 5). Results show selection bias mitigating, when adjusted with informed smoothing with $k > 10$, can increase accuracy over baseline.

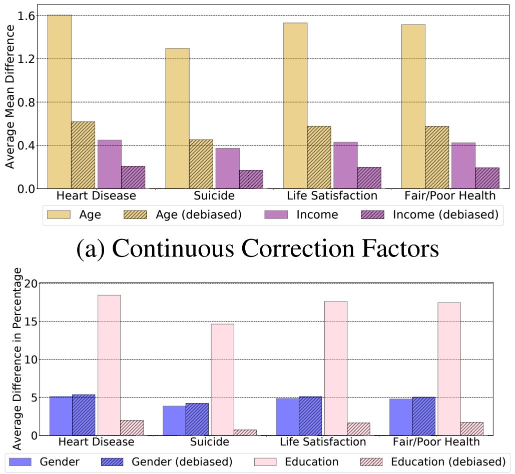
a slight increase using age and larger increases for income and education. Consistent with our previous results, we see no improvements for gender correction. All results converge to no correction with large enough k since the Informed Smoothing has the effect of backing off to assuming the county is fully representative (i.e. no correction).[6]

Figures 3(b) and 3(c) show Informed Smoothing with naive and raking factors, respectively. We see that the combination of age and gender does not drastically improve over baseline. We also see raking helping more for the age-gender-income-education correction factor (for both Informed Smoothing and Adaptive Binning), suggesting that raking might work better than naive post-stratification as the number of correction factors increases.

Recommended System Due to the large number of tuning parameters evaluated above, we perform a backwards elimination on our correction variables to find the model with

Baseline = .640	Minimum Count Threshold				
	1	10	50	100	1000
Age	.583	.605$^+$	.624$^+$	.636$^+$	.634$^+$
Gender	.639	.639	.639	.640	.640
Income	.612$^-$	.666*	.674*	.663*	.642$^+$
Education	.648*	.648*	.648*	.647*	.642
Age + Gender					
Naive	.580	.598$^+$	.622$^+$	.633$^+$	.633$^+$
Raking	.580	.603$^+$	.623$^+$	.635$^+$	.634$^+$
Inc. + Edu.					
Naive	.612$^-$	.659*	.673*	.662*	.643
Raking	.611$^-$	.662*	.674*	.664*	.643
All					
Naive	.634	.633	.620$^-$	.634	.645$^+$
Raking	.579$^-$	.610$^+$	.634$^+$	.649*	.647*

Table 6: Average predictive accuracies (Pearson r) across four tasks when using *adaptive binning*. $^+$ and $^-$ indicate a significant increase or decrease, respectively, as compared to the same correction variable / method pair in Table 5, * increase over baseline. "All" includes age, gender, income, and education. This method shows mitigating selection bias can improve predictive accuracy when adjusting for error in demographic scores by using adaptive binning.



(a) Continuous Correction Factors

(b) Categorical Correction Factors

Figure 4: Effect of mitigation on average bias. Bars indicate bias for both (a) continuous attributes (age and income) quantified as difference in standardized means, or (b) dichotomous attributes (gender and education) quantified as percentage difference between census populations and our measurements. Darker bars indicate the debiased version, from our final suggested approach using estimator redistribution, adaptive binning, and informed smoothing – settings from the last line of Table 7). Bias is reduced for all factors except gender (which only had a small baseline bias).

[6] Uninformed smoothing, such as Lapacian smoothing, would push counties toward a non-representative (uniform) distribution, negating the point of selection bias correction. See Supplement[2] for an "add one" smoothing.

	Baseline			Optimal Model based on Backwards Elimination					
	Pearson r	R^2	*RMSE*	Strat. Vars.	Adaptive Binning	Smoothing k	Pearson r	R^2	*RMSE*
Heart Disease	.753	.563	*30.14*	inc. + edu.	50	10	.769*	.588	*29.26*
Suicide	.614	.371	*3.67*	inc.	50	10	.626*	.388	*3.62*
Life Satisfaction	.445	.187	*0.024*	inc.	50	10	.542*	.286	*0.022*
Poor/Fair Health	.746	.552	*3.82*	inc. + edu.	50	10	.778*	.603	*3.60*
Average	.640	.418	*9.41*	inc.	50	10	.677*	.464	*9.18*
				inc + edu	50	10	.676*	.463	*9.13*

Table 7: Predictive performance of baseline vs. our recommended system. Across each of our four individual county-level prediction tasks we present a recommended system (i.e., the combination of settings which resulted in the highest prediction accuracy with the smallest number of correction factors). * indicates significant reduction in error over baseline; $p < .05$.

the highest accuracy and the smallest number of correction factors. To make this selection process simpler, we evaluate our backwards selection using raking, as this method tended to outperform naive post-stratification, as well as estimator redistribution. The algorithm behaves as follows. First, we start with the maximum number of correction factors (age, gender, income, and education) and then perform a grid search over Adaptive Binning and Informed Smoothing parameters. We then choose a smaller set of correction factors and perform the same grid search, until we find a model with the smallest number of correction factors that gives us the best accuracy. This evaluation is performed across our four county-level outcomes, as opposed to an average of the four. Table 7 shows predictive accuracies for each of our four outcomes using the above algorithm. Here we see a fairly consistent set of tuning parameters: a minimum bin threshold of 50, a smoothing constant $k = 10$, and income as a correction factor. For two of the our outcomes, heart disease and poor/fair health, we see education as an additional correction factor. The maximum percent increase occurs for life satisfaction (52.9%) while heart disease has the smallest significant increase (4.44%). In the end, both models are very similar, with the only difference being the addition of education correction. Thus, in the end, we recommend using both income and education since, everything else being equal, we believe it is better to correct for more factors in order to make the model more "fair". In Figure 5 show county level heat maps of the true values of the Poor/Fair Health measure (as reported by the BRFSS) and Twitter predicted values (as determined by cross validation using our recommended system). Here we see our recommended system in Figure 5(a) reasonably tracking the ground truth values in Figure 5(b).

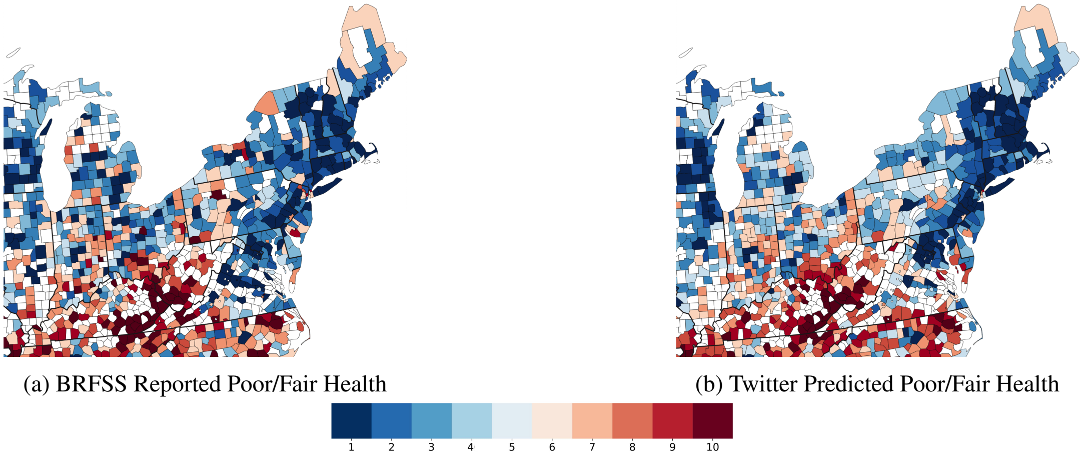
Quantifying Bias Reductions

Here we quantify the reduction in bias in our recommended system. For our two continuous variables (age and income) we take the absolute difference in means of the census bin percentages and our estimated demographics, normalized by the pooled standard deviation, and average across all counties (i.e., average absolute Cohen's D). We do this once for our "out-of-the-box" socio-demographic estimates (to get a baseline bias) and again for the weighted estimates from our recommended model (i.e., correcting for income and education with a minimum bin threshold of 50 and smoothing $k = 10$). For the two categorical variables (gender and education), we take the absolute difference in county percentage female and percentage with a bachelor's degree, and then average across all counties. This process is repeated for each of our four county level tasks.

Figure 4 shows the results of this experiment. Here we see a significant reduction in bias for age, income, and education across all four tasks. Gender, on the other hand, shows a slight increase in bias, though we note that this increase is not nearly as dramatic as the decreases seen across age, income, and education, nor have we attempted to fully correct for gender (as this model corrects for income and education only). These results seem to match the county-level statistics reported in Table 2, which shows a small difference difference in gender between the Census data and our Twitter sample. That is, we do not expect our methods to drastically address gender biases because gender is evenly distributed across counties, in both the Census data and our Twitter sample. Finally, we note that, while our recommended system only performs raking across income and education, we apply estimator redistribution across all four correction factors (age, gender, income, and education). Thus, we might expect our final bias measures to correct for age despite the fact our post-stratification process does not consider this variable.

Conclusion

Selection bias — producing measurements over a population sample that differs from the target population — is a frequent criticism of automatic social media-based population predictions (Hoover and Dehghani 2020; Wang et al. 2019; Shah, Schwartz, and Hovy 2020). While post-stratification techniques are frequently used to address selection bias in opinion polling or social sciences, we found "out of the box" methods generally resulted in worse performance, as compared to no correction, for the task of predicting population (i.e., U.S. county) health and well-being statistics from social media language. We discovered two reasons for this lack of benefit: (1) *estimating* sample user demographics from predictive models (as opposed to having self-reported demographics of the Twitter users) introduces additional biases when compared to known distributions and (2) sparse or underpowered data for estimating the observed community demographic distribu-



(a) BRFSS Reported Poor/Fair Health

(b) Twitter Predicted Poor/Fair Health

Figure 5: Maps of northeastern U.S. counties showing the deciles of adults living in Poor/Fair Health (a) as reported by the BRFSS and (b) as predicted from Twitter, using our recommended model in Table 7. Out-of-sample Twitter predictions obtained though the cross validation process. Blue is less poor/fair health (i.e., better health); red is more poor/fair health (i.e., worse health), and white is unreliable self-report or Twitter data (i.e., missing data).

tions. To the best of our knowledge, neither of these issues has been previously investigated for improving post-stratification. In fact, few works have even evaluated commonly used selection bias mitigation techniques for predictive tasks (Wang et al. 2019; Culotta 2014), likely because such techniques are traditionally applied without access to ground truth validation data (e.g., in most opinion polls).

We proposed ROBUST POSTSTRATIFICATION which includes several techniques to address challenges in selection bias correction for predicting population statistics and evaluating their efficacy. First, we found that using *estimator redistribution* to counter shrinkage bias of estimated demographics provided modest benefits. Then, we explored two techniques for addressing sparse bin issues: *adaptive binning* and *informed smoothing*, finding both provided a substantial benefit and resulted in an overall improvement to the predictive models, yielding state-of-the-art results (a 52.9% increase in variance explained for life satisfaction). Many approaches for addressing demographic biases in AI try to correct them without sacrificing accuracy (Gonen and Goldberg 2019). In the case of selection bias, we believe we have shown that properly correcting bias can yield substantial benefits.

References

Abebe, R.; Giorgi, S.; Tedijanto, A.; Buffone, A.; and Schwartz, H. A. 2020. Quantifying Community Characteristics of Maternal Mortality Using Social Media. In *The World Wide Web Conference*.

Alexander, C. H. 1987. A class of methods for using person controls in household weighting. *Survey Methodology*, 13(2): 183–198.

Battaglia, M. P.; Izrael, D.; Hoaglin, D. C.; and Frankel, M. R. 2009. Practical considerations in raking survey data. *Survey Practice*, 2(5): 1–10.

Berk, R. A.; and Ray, S. C. 1982. Selection biases in sociological data. *Social Science Research*, 11(4): 352–398.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Blodgett, S. L.; Green, L.; and O'Connor, B. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1119–1130. Austin, Texas: Association for Computational Linguistics.

Chen, X.; Wang, Y.; Agichtein, E.; and Wang, F. 2015. A comparative study of demographic attribute inference in twitter. In *ICWSM*, 590–593. Oxford, UK: The AAAI Press.

Coppersmith, G.; Dredze, M.; Harman, C.; and Hollingshead, K. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1–10. Denver, Colorado: Association for Computational Linguistics.

Culotta, A. 2014. Reducing sampling bias in social media data for county health inference. In *Joint Statistical Meetings Proceedings*, 1–12. Boston, MA: American Statistical Association.

Curtis, B.; Giorgi, S.; Buffone, A. E.; Ungar, L. H.; Ashford, R. D.; Hemmons, J.; Summers, D.; Hamilton, C.; and Schwartz, H. A. 2018. Can Twitter be used to predict county excessive alcohol consumption rates? *PloS one*, 13(4): e0194290.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 47–56. ACM.

Deville, J.-C.; Särndal, C.-E.; and Sautory, O. 1993. Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423): 1013–1020.

Duggan, M.; and Smith, A. 2013. Demographics of Key Social Networking Platforms. *Pew Research*.

Eichstaedt, J. C.; Schwartz, H. A.; Kern, M. L.; Park, G.; Labarthe, D. R.; Merchant, R. M.; Jha, S.; Agrawal, M.; Dziurzynski, L. A.; Sap, M.; et al. 2015. Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological science*, 0956797614557867.

Giorgi, S.; Preoţiuc-Pietro, D.; Buffone, A.; Rieman, D.; Ungar, L.; and Schwartz, H. A. 2018. The Remarkable Benefit of User-Level Aggregation for Lexical-based Population-Level Predictions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1167–1172. Brussels, Belgium: Association for Computational Linguistics.

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614.

Greenwood, S.; Perrin, A.; and Duggan, M. 2016. Social media update 2016: Facebook usage and engagement is on the rise, while adoption of other platforms holds steady. *Pew Research Center*.

Hecht, B. J.; and Stephens, M. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. *ICWSM*, 14: 197–205.

Henry, K.; and Valliant, R. 2012. Methods for adjusting survey weights when estimating a total. *Proceedings of the Federal Committee on Statistical Methodology, January*, 10–12.

Hoerl, A. E.; and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1): 55–67.

Holt, D.; and Smith, T. F. 1979. Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, 142(1): 33–46.

Hoover, J.; and Dehghani, M. 2020. The big, the bad, and the ugly: Geographic estimation with flawed psychological data. *Psychological Methods*, 25(4): 412.

Jaidka, K.; Giorgi, S.; Schwartz, H. A.; Kern, M. L.; Ungar, L. H.; and Eichstaedt, J. C. 2020. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*.

Kalton, G.; and Flores-Cervantes, I. 2003. Weighting methods. *Journal of official statistics*, 19(2): 81.

Kneser, R.; and Ney, H. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, 181–184. IEEE.

Kost, J. T.; and McDermott, M. P. 2002. Combining dependent P-values. *Statistics & Probability Letters*, 60(2): 183–190.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Lawless, N. M.; and Lucas, R. E. 2011. Predictors of regional well-being: A county level analysis. *Social Indicators Research*, 101(3): 341–357.

Leemann, L.; and Wasserfallen, F. 2017. Extending the use and prediction precision of subnational public opinion estimation. *American journal of political science*, 61(4): 1003–1022.

Little, R. J. 1993. Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, 88(423): 1001–1012.

Lui, M.; and Baldwin, T. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, 25–30. Jeju Island, Korea: Association for Computational Linguistics.

Matz, S. C.; Menges, J. I.; Stillwell, D. J.; and Schwartz, H. A. 2019. Predicting individual-level income from Facebook profiles. *PloS one*.

Miranda Filho, R.; Almeida, J. M.; and Pappa, G. L. 2015. Twitter population sample bias and its impact on predictive outcomes: a case study on elections. In *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, 1254–1261. IEEE.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the Demographics of Twitter Users. *ICWSM*, 11: 5th.

Mowery, D. L.; Park, A.; Bryan, C.; and Conway, M. 2016. Towards Automatically Classifying Depressive Symptoms from Twitter Data for Population Health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, 182–191. Osaka, Japan: The COLING 2016 Organizing Committee.

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129): 1–2.

Park, D. K.; Gelman, A.; and Bafumi, J. 2004. Bayesian multilevel estimation with poststratification: State-level estimates from national polls. *Political Analysis*, 12(4): 375–385.

Pavalanathan, U.; and Eisenstein, J. 2015. Confounds and consequences in geotagged Twitter data. *EMNLP*.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, 729–740.

Remington, P. L.; Catlin, B. B.; and Gennuso, K. P. 2015. The county health rankings: rationale and methods. *Population health metrics*, 13(1): 11.

Sap, M.; Park, G.; Eichstaedt, J.; Kern, M.; Stillwell, D.; Kosinski, M.; Ungar, L.; and Schwartz, H. A. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), 1146–1151. Doha, Qatar: Association for Computational Linguistics.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Agrawal, M.; Park, G. J.; Lakshmikanth, S. K.; Jha, S.; Seligman, M. E.; Ungar, L.; et al. 2013a. Characterizing geographic variation in well-being using tweets. In *ICWSM*.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013b. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9): e73791.

Schwartz, H. A.; Giorgi, S.; Sap, M.; Crutchley, P.; Eichstaedt, J. C.; and Ungar, L. 2017. DLATK: Differential Language Analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP.

Shah, D. S.; Schwartz, H. A.; and Hovy, D. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264.

Wang, Z.; Hale, S.; Adelani, D. I.; Grabowicz, P.; Hartman, T.; Jurgens, D.; et al. 2019. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In *The World Wide Web Conference*, 2056–2067. ACM.

Weeg, C.; Schwartz, H. A.; Hill, S.; Merchant, R. M.; Arango, C.; and Ungar, L. 2015. Using Twitter to measure public discussion of diseases: a case study. *JMIR public health and surveillance*, 1(1).

Winship, C.; and Mare, R. D. 1992. Models for sample selection bias. *Annual review of sociology*, 18(1): 327–350.

Zagheni, E.; and Weber, I. 2015. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1): 13–25.

Zhang, J.; Hu, X.; Zhang, Y.; and Liu, H. 2016. Your age is no secret: Inferring microbloggers' ages via content and interaction analysis. In *ICWSM*. Cologne, Germany: The AAAI Press.