Confounds and Consequences in Geotagged Twitter Data

Umashanthi Pavalanathan and **Jacob Eisenstein**
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30308
{umashanthi + jacobe}@gatech.edu

Abstract

Twitter is often used in quantitative studies that identify geographically-preferred topics, writing styles, and entities. These studies rely on either GPS coordinates attached to individual messages, or on the user-supplied location field in each profile. In this paper, we compare these data acquisition techniques and quantify the biases that they introduce; we also measure their effects on linguistic analysis and text-based geolocation. GPS-tagging and self-reported locations yield measurably different corpora, and these linguistic differences are partially attributable to differences in dataset composition by age and gender. Using a latent variable model to induce age and gender, we show how these demographic variables interact with geography to affect language use. We also show that the accuracy of text-based geolocation varies with population demographics, giving the best results for men above the age of 40.

1 Introduction

Social media data such as Twitter is frequently used to identify the unique characteristics of geographical regions, including topics of interest (Hong et al., 2012), linguistic styles and dialects (Eisenstein et al., 2010; Gonçalves and Sánchez, 2014), political opinions (Caldarelli et al., 2014), and public health (Broniatowski et al., 2013). Social media permits the aggregation of datasets that are orders of magnitude larger than could be assembled via traditional survey techniques, enabling analysis that is simultaneously fine-grained and global in scale. Yet social media is not a representative sample of any "real world" population, aside from social media itself. Using social media as a sample therefore risks introducing both geographic and demographic biases (Mislove et al., 2011; Hecht and Stephens, 2014; Longley et al., 2015; Malik et al., 2015).

This paper examines the effects of these biases on the geo-linguistic inferences that can be drawn from Twitter. We focus on the ten largest metropolitan areas in the United States, and consider three sampling techniques: drawing an equal number of GPS-tagged tweets from each area; drawing a *county-balanced* sample of GPS-tagged messages to correct Twitter's urban skew (Hecht and Stephens, 2014); and drawing a sample of *location-annotated* messages, using the location field in the user profile. Leveraging self-reported first names and census statistics, we show that the age and gender composition of these datasets differ significantly.

Next, we apply standard methods from the literature to identify geo-linguistic differences, and test how the outcomes of these methods depend on the sampling technique and on the underlying demographics. We also test the accuracy of text-based geolocation (Cheng et al., 2010; Eisenstein et al., 2010) in each dataset, to determine whether the accuracies reported in recent work will generalize to more balanced samples.

The paper reports several new findings about geotagged Twitter data:

- In comparison with tweets with self-reported locations, GPS-tagged tweets are written more often by young people and by women.
- There are corresponding linguistic differences between these datasets, with GPS-tagged tweets including more geographically-specific non-standard words.
- Young people use significantly more geographically-specific non-standard words. Men tend to mention more geographically-specific entities than women, but these

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2138–2148,
Lisbon, Portugal, 17-21 September 2015.

differences are significant only for individuals at the age of 30 or older.

- Users who GPS-tag their tweets tend to write more, making them easier to geolocate. Evaluating text-based geolocation on GPS-tagged tweets probably overestimates its accuracy.
- Text-based geolocation is significantly more accurate for men and for older people.

These findings should inform future attempts to generalize from geotagged Twitter data, and may suggest investigations into the demographic properties of other social media sites.

We first describe the basic data collection principles that hold throughout the paper (§ 2). The following three sections tackle demographic biases (§ 3), their linguistic consequences (§ 4), and the impact on text-based geolocation (§ 5); each of these sections begins with a discussion of methods, and then presents results. We then summarize related work and conclude.

2 Dataset

This study is performed on a dataset of tweets gathered from Twitter's streaming API from February 2014 to January 2015. During an initial filtering step we removed retweets, repetitions of previously posted messages which contain the "retweeted_status" metadata or "RT" token which is widely used among Twitter users to indicate a retweet. To eliminate spam and automated accounts (Yardi et al., 2009), we removed tweets containing URLs, user accounts with more than 1000 followers or followees, accounts which have tweeted more than 5000 messages at the time of data collection, and the top 10% of accounts based on number of messages in our dataset. We also removed users who have written more than 10% of their tweets in any language other than English, using Twitter's `lang` metadata field. Exploration of code-switching (Solorio and Liu, 2008) and the role of second-language English speakers (Eleta and Golbeck, 2014) is left for future work.

We consider the ten largest Metropolitan Statistical Areas (MSAs) in the United States, listed in Table 1. MSAs are defined by the U.S. Census Bureau as geographical regions of high population with density organized around a single urban core; they are not legal administrative divisions. MSAs include outlying areas that may be substantially less urban than the core itself. For example, the Atlanta MSA is centered on Fulton County (1750 people per square mile), but extends to Haralson County (100 people per square mile), on the border of Alabama. A per-county analysis of this data therefore enables us to assess the degree to which Twitter's skew towards urban areas biases geo-linguistic analysis.

3 Representativeness of geotagged Twitter data

We first assess potential biases in sampling techniques for obtaining geotagged Twitter data. In particular, we compare two possible techniques for obtaining data: the location field in the user profile (Poblete et al., 2011; Dredze et al., 2013), and the GPS coordinates attached to each message (Cheng et al., 2010; Eisenstein et al., 2010).

3.1 Methods

To build a dataset of GPS-tagged messages, we extracted the GPS latitude and longitude coordinates reported in the tweet, and used GIS-TOOLS[1] reverse geocoding to identify the corresponding counties. This set of geotagged messages will be denoted $\mathcal{D}_G$. Only 1.24% of messages contain geo-coordinates, and it is possible that the individuals willing to share their GPS comprise a skewed population. We therefore also considered the user-reported location field in the Twitter profile, focusing on the two most widely-used patterns: (1) city name, (2) city name and two letter state name (e.g. *Chicago* and *Chicago, IL*). Messages that matched any of the ten largest MSAs were grouped into a second set, $\mathcal{D}_L$.

While the inconsistencies of writing style in the Twitter location field are well-known (Hecht et al., 2011), analysis of the intersection between $\mathcal{D}_G$ and $\mathcal{D}_L$ found that the two data sources agreed the overwhelming majority of the time, suggesting that most self-provided locations are accurate. Of course, there may be many false negatives — profiles that we fail to geolocate due to the use of non-standard toponyms like *Pixburgh* and *ATL*. If so, this would introduce a bias in the population sample in $\mathcal{D}_L$. Such a bias might have linguistic consequences, with datasets based on the location field containing less non-standard language overall.

[1] `https://github.com/DrSkippy/Data-Science-45min-Intros/blob/master/gis-tools-101/gis_tools.ipynb`

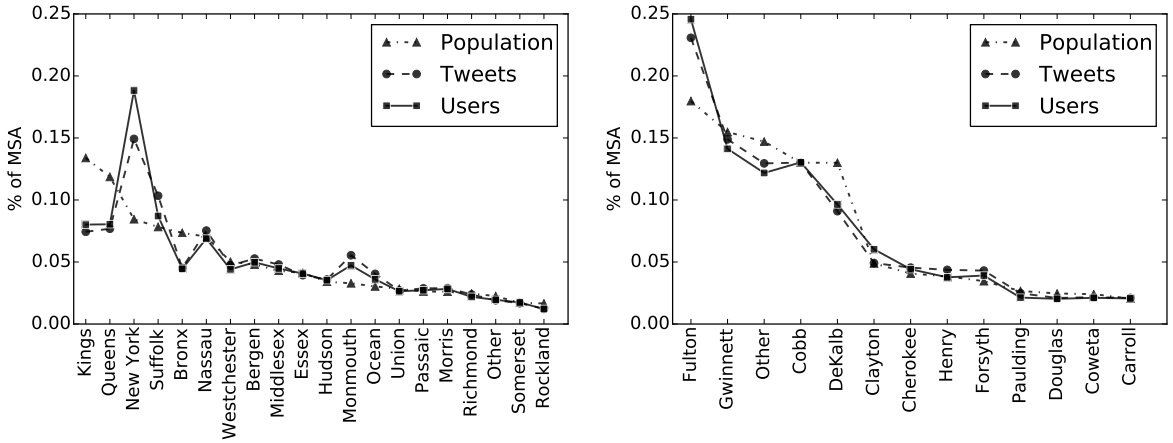


Figure 1: Proportion of census population, Twitter messages, and Twitter user accounts, by county. New York is shown on the left, Atlanta on the right.

3.1.1 Subsampling

The initial samples $\mathcal{D}_G$ and $\mathcal{D}_L$ were then resampled to create the following balanced datasets:

GPS-MSA-BALANCED From $\mathcal{D}_G$, we randomly sampled 25,000 tweets per MSA as the message-balanced sample, and all the tweets from 2,500 users per MSA as the user-balanced sample. Balancing across MSAs ensures that the largest MSAs do not dominate the linguistic analysis.

GPS-COUNTY-BALANCED We resampled $\mathcal{D}_G$ based on county-level population (obtained from the U.S. Census Bureau), and again obtained message-balanced and user-balanced samples. These samples are more geographically representative of the overall population distribution across each MSA.

LOC-MSA-BALANCED From $\mathcal{D}_L$, we randomly sampled 25,000 tweets per MSA as the message-balanced sample, and all the tweets from 2,500 users per MSA as the user-balanced sample. It is not possible to obtain county-level geolocations in $\mathcal{D}_L$, as exact geographical coordinates are unavailable.

3.1.2 Age and gender identification

To estimate the distribution of ages and genders in each sample, we queried statistics from the Social Security Administration, which records the number of individuals born each year with each given name. Using this information, we obtained the probability distribution of age values for each given name. We then matched the names against the first token in the name field of each user's profile, enabling us to induce approximate distributions over ages and genders. Unlike Facebook and Google+, Twitter does not have a "real name" policy, so users are free to give names that are fake, humorous, etc. We eliminate user accounts whose names are not sufficiently common in the social security database (i.e. first names which are at least 100 times more frequent in Twitter than in the social security database), thereby omitting 33% of user accounts, and 34% of tweets. While some individuals will choose names not typically associated with their gender, we assume that this will happen with roughly equal probability in both directions. So, with these caveats in mind, we induce the age distribution for the GPS-MSA-BALANCED sample and the LOC-MSA-BALANCED sample as,

$$p(a \mid \text{name} = n) = \frac{count(\text{name} = n, \text{age} = a)}{\sum_{a'} count(\text{name} = n, \text{age} = a')} \quad (1)$$

$$p_{\mathcal{D}}(a) \propto \sum_{i \in \mathcal{D}} p(a \mid \text{name} = n_i). \quad (2)$$

We induce distributions over author gender in much the same way (Mislove et al., 2011). This method does not incorporate prior information about the ages of Twitter users, and thus assigns too much probability to the extremely young and old, who are unlikely to use the service. While it would be easy to design such a prior — for example, assigning zero prior probability to users under the age of five or above the age of 95 — we see no principled basis for determining these cutoffs. We therefore focus on the *differences* between the estimated $p_{\mathcal{D}}(a)$ for each sample $\mathcal{D}$.

MSA	Num. Counties	L1 Dist. Population vs. Users	L1 Dist. Population vs. Tweets
New York	23	0.2891	0.2825
Los Angeles	2	0.0203	0.0223
Chicago	14	0.0482	0.0535
Dallas	12	0.1437	0.1176
Houston	10	0.0394	0.0472
Philadelphia	11	0.1426	0.1202
Washington DC	22	0.2089	0.2750
Miami	3	0.0428	0.0362
Atlanta	28	0.1448	0.1730
Boston	7	0.1878	0.2303

Table 1: L1 distance between county-level population and Twitter users and messages

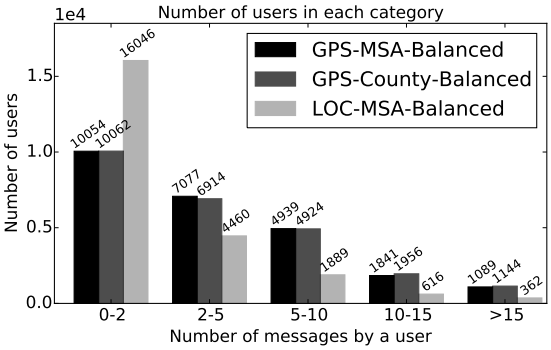


Figure 2: User counts by number of Twitter messages

3.2 Results

Geographical biases in the GPS Sample We first assess the differences between the true population distributions over counties, and the per-tweet and per-user distributions. Because counties vary widely in their degree of urbanization and other demographic characteristics, this measure is a proxy for the representativeness of GPS-based Twitter samples (county information is not available for the LOC-MSA-BALANCED sample). Population distributions for New York and Atlanta are shown in Figure 1. In Atlanta, Fulton County is the most populous and most urban, and is over-represented in both geotagged tweets and user accounts; most of the remaining counties are correspondingly underrepresented. This coheres with the urban bias noted earlier by Hecht and Stephens (2014). In New York, Kings County (Brooklyn) is the most populous, but is underrepresented in both the number of geotagged tweets and user accounts, at the expense of New York County (Manhattan). Manhattan is the commercial and entertainment center of the New York MSA, so residents of outlying counties may be tweeting from their jobs or social activities.

To quantify the representativeness of each sample, we use the L1 distance $||\mathbf{x} - \mathbf{y}||_1 = \sum_c |p_c - t_c|$, where p_c is the proportion of the MSA population residing in county c and t_c is the proportion of tweets (Table 1). County boundaries are determined by states, and their density varies: for example, the Los Angeles MSA covers only two counties, while the smaller Atlanta MSA is spread over 28 counties. The table shows that while New York is the most extreme example, most MSAs feature an asymmetry between county population and Twitter adoption.

Usage Next, we turn to differences between the GPS-based and profile-based techniques for obtaining ground truth data. As shown in Figure 2, the LOC-MSA-BALANCED sample contains more low-volume users than either the GPS-MSA-BALANCED or GPS-COUNTY-BALANCED samples. We can therefore conclude that the county-level geographical bias in the GPS-based data does not impact usage rate, but that the difference between GPS-based and profile-based sampling does; the linguistic consequences of this difference will be explored in the following sections.

Demographics Table 2 shows the expected age and gender for each dataset, with bootstrap confidence intervals. Users in the LOC-MSA-BALANCED dataset are on average two years older than in the GPS-MSA-BALANCED and GPS-COUNTY-BALANCED datasets, which are statistically indistinguishable. Focusing on the difference between GPS-MSA-BALANCED and LOC-MSA-BALANCED, we plot the difference in age probabilities in Figure 3, showing that GPS-MSA-BALANCED includes many more teens and people in their early twenties, while LOC-MSA-BALANCED includes more people at middle age and older. Young people are especially likely to use social media on cellphones (Lenhart, 2015), where location tagging would be more relevant than when Twitter is accessed via a personal computer. Social media users in the age brackets 18-29 and 30-49 are also more likely to tag their locations in social media posts than social media users in the age brackets 50-64 and 65+ (Zickuhr, 2013), with women and men tagging at roughly equal rates. Table 2 shows that the GPS-MSA-BALANCED and GPS-COUNTY-BALANCED samples contain significantly more women than LOC-

Sample	Expected Age	95% CI	% Female	95% CI
GPS-MSA-BALANCED	36.17	[36.07 – 36.27]	51.5	[51.3 – 51.8]
GPS-COUNTY-BALANCED	36.25	[36.16 – 36.30]	51.3	[51.1 – 51.6]
LOC-MSA-BALANCED	38.35	[38.25 – 38.44]	49.3	[49.1 – 49.6]

Table 2: Demographic statistics for each dataset

Figure 3: Difference in age probability distributions between GPS-MSA-BALANCED and LOC-MSA-BALANCED.

MSA-BALANCED, though all three samples are close to 50%.

4 Impact on linguistic generalizations

Many papers use Twitter data to draw conclusions about the relationship between language and geography. What role do the demographic differences identified in the previous section have on the linguistic conclusions that emerge? We measure the differences between the linguistic corpora obtained by each data acquisition approach. Since the GPS-MSA-BALANCED and GPS-COUNTY-BALANCED methods have nearly identical patterns of usage and demographics, we focus on the difference between GPS-MSA-BALANCED and LOC-MSA-BALANCED. These datasets differ in age and gender, so we also directly measure the impact of these demographic factors on the use of geographically-specific linguistic variables.

4.1 Methods

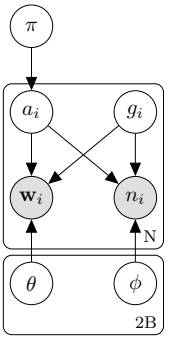
Discovering geographical linguistic variables We focus on lexical variation, which is relatively easy to identify in text corpora. Monroe et al. (2008) survey a range of alternative statistics for finding lexical variables, demonstrating that a regularized log-odds ratio strikes a good balance between distinctiveness and robustness. A similar approach is implemented in SAGE (Eisenstein et al., 2011a)[2], which we use here. For each sample — GPS-MSA-BALANCED and LOC-MSA-BALANCED — we apply SAGE to identify the twenty-five most salient lexical items for each metropolitan area.

Keyword annotation Previous research has identified two main types of geographical lexical variables. The first are non-standard words and spellings, such as *hella* and *yinz*, which have been found to be very frequent in social media (Eisenstein, 2015). Other researchers have focused on the "long tail" of entity names (Roller et al., 2012). A key question is the relative importance of these two variable types, since this would decide whether geo-linguistic differences are primarily topic-based or stylistic. It is therefore important to know whether the frequency of these two variable types depends on properties of the sample. To test this, we take the lexical items identified by SAGE (25 per MSA, for both the GPS-MSA-BALANCED and LOC-MSA-BALANCED samples), and annotate them as NONSTANDARD-WORD, ENTITY-NAME, or OTHER. Annotation for ambiguous cases is based on the majority sense in randomly-selected examples. Overall, we identify 24 NONSTANDARD-WORDS and 185 ENTITY-NAMES.

Inferring author demographics As described in § 3.1.2, we can obtain an approximate distribution over author age and gender by linking self-reported first names with aggregate statistics from the United States Census. To sharpen these estimates, we now consider the text as well, building a simple latent variable model in which both the name and the word counts are drawn from distributions associated with the latent age and gender (Chang et al., 2010). The model is shown in Figure 4, and involves the following generative process:

For each user $i \in \{1...N\}$,
(a) draw the age, $a_i \sim \text{Categorical}(\pi)$
(b) draw the gender, $g_i \sim \text{Categorical}(0.5)$

[2] https://github.com/jacobeisenstein/jos-gender-2014



a_i	Age (bin) for author i
g_i	Gender of author i
$\mathbf{w}_i$	Word counts for author i
n_i	First name of author i
π	Prior distribution over age bins
$\theta_{a,g}$	Word distribution for age a and gender g
$\phi_{a,g}$	First name distribution for age a and gender g

Figure 4: Plate diagram for latent variable model of age and gender

(c) draw the author's given name, $n_i \sim \text{Categorical}(\phi_{a_i,g_i})$

(d) draw the word counts, $w_i \sim \text{Multinomial}(\theta_{a_i,g_i})$,

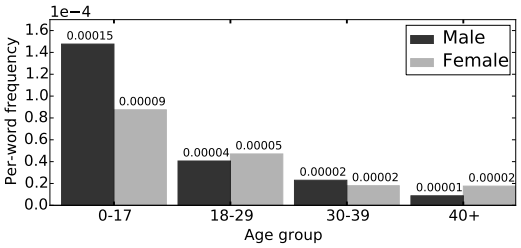
where we elide the second parameter of the multinomial distribution, the total word count. We use expectation-maximization to perform inference in this model, binning the latent age variable into four groups: 0-17, 18-29, 30-39, above 40.[3] Because the distribution of names given demographics is available from the Social Security data, we clamp the value of ϕ throughout the EM procedure. Other work in the domain of demographic prediction often involves more complex methods (Nguyen et al., 2014; Volkova and Durme, 2015), but since it is not the focus of our research, we take a relatively simple approach here, assuming no labeled data for demographic attributes.

4.2 Results

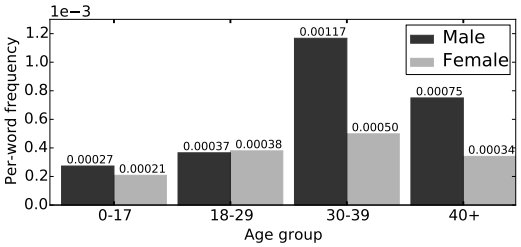
Linguistic differences by dataset We first consider the impact of the data acquisition technique on the lexical features associated with each city. The keywords identified in GPS-MSA-BALANCED dataset feature more geographically-specific non-standard words, which occur at a rate of 3.9×10^{-4} in GPS-MSA-BALANCED, versus 2.6×10^{-4} in LOC-MSA-BALANCED; this difference is statistically significant ($p < .05, t = 3.2$).[4]

[3] Binning is often employed in work on text-based age prediction (Garera and Yarowsky, 2009; Rao et al., 2010; Rosenthal and McKeown, 2011); it enables word and name counts to be shared over multiple ages, and avoids the complexity inherent in regressing a high-dimensional textual predictors against a numerical variable.

[4] We employ a paired t-test, comparing the difference in frequency for each word across the two datasets. Since we cannot test the complete set of entity names or non-standard words, this quantifies whether the observed difference is robust across the subset of the vocabulary that we have selected.



(a) non-standard words



(b) entity names

Figure 5: Aggregate statistics for geographically-specific non-standard words and entity names across imputed demographic groups, from the GPS-MSA-BALANCED sample.

For entity names, the difference between datasets was not significant, with a rate of 4.0×10^{-3} for GPS-MSA-BALANCED, and 3.7×10^{-3} for LOC-MSA-BALANCED. Note that these rates include only the non-standard words and entity names detected by SAGE as among the top 25 most distinctive for one of the ten largest cities in the US; of course there are many other relevant terms that are below this threshold.

In a pilot study of the GPS-COUNTY-BALANCED data, we found few linguistic differences from GPS-MSA-BALANCED, in either the aggregate word-group frequencies or the SAGE word lists — despite the geographical imbalances shown in Table 1 and Figure 1. Informal examination of specific counties shows some expected differences: for example, Clayton County, which hosts Atlanta's Hartsfield-Jackson airport, includes terms related to air travel, and other counties include mentions of local cities and business districts. But the aggregate statistics for under-represented counties are not substantially different from those of overrepresented counties, and are largely unaffected by county-based resampling.

Demographics Aggregate linguistic statistics for demographic groups are shown in Figure 5. Men use significantly more geographically-specific entity names than women ($p \ll .01, t =$

Age	Sex	New York	Dallas
0-17	F	*niall, ilysm, hemmings, stalk, ily*	*fanuary, idol, lmbo, lowkey, jonas*
	M	*ight, technique, kisses, lesbian, dicks*	*homies, daniels, oomf, teenager, brah*
18-29	F	*roses, castle, hmmmm, chem, sinking*	*socially, coma, hubby, bra, swimming*
	M	*drunken, manhattan, spoiler, guardians, gonna*	*harden, watt, astros, rockets, mavs*
30-39	F	*suite, nyc, colleagues, york, portugal*	*astros, sophia, recommendations, houston, prepping*
	M	*mets, effectively, cruz, founder, knicks*	*texans, rockets, embarrassment, tcu, mississippi*
40+	F	*cultural, affected, encouraged, proverb, unhappy*	*determine, islam, rejoice, psalm, responsibility*
	M	*reuters, investors, shares, lawsuit, theaters*	*mph, wazers, houston, tx, harris*

Table 3: Most characteristic words for demographic subsets of each city, as compared with the overall average word distribution

8.0), but gender differences for geographically-specific non-standard words are not significant ($p \approx .2$).[5] Younger people use significantly more geographically-specific non-standard words than older people (ages 0–29 versus 30+, $p \ll .01, t = 7.8$), and older people mention significantly more geographically-specific entity names ($p \ll .01, t = 5.1$). Of particular interest is the intersection of age and gender: the use of geographically-specific non-standard words decreases with age much more profoundly for men than for women; conversely, the frequency of mentioning geographically-specific entity names increases dramatically with age for men, but to a much lesser extent for women. The observation that high-level patterns of geographically-oriented language are more age-dependent for men than for women suggests an intriguing site for future research on the intersectional construction of linguistic identity.

For a more detailed view, we apply SAGE to identify the most salient lexical items for each MSA, subgrouped by age and gender. Table 3 shows word lists for New York (the largest MSA) and Dallas (the 5th-largest MSA), using the GPS-MSA-BALANCED sample. Non-standard words tend to be used by the youngest authors: *ilysm* ('I love you so much'), *ight* ('alright'), *oomf* ('one of my followers'). Older authors write more about local entities (*manhattan, nyc, houston*), with men focusing on sports-related entities (*harden, watt, astros, mets, texans*), and women above the age of 40 emphasizing religiously-oriented terms (*proverb, islam, rejoice, psalm*).

[5]But see Bamman et al. (2014) for a much more detailed discussion of gender and standardness.

5 Impact on text-based geolocation

A major application of geotagged social media is to predict the geolocation of individuals based on their text (Eisenstein et al., 2010; Cheng et al., 2010; Wing and Baldridge, 2011; Hong et al., 2012; Han et al., 2014). Text-based geolocation has obvious commercial implications for location-based marketing and opinion analysis; it is also potentially useful for researchers who want to measure geographical phenomena in social media, and wish to access a larger set of individuals than those who provide their locations explicitly.

Previous research has obtained impressive accuracies for text-based geolocation: for example, Hong et al. (2012) report a median error of 120 km, which is roughly the distance from Los Angeles to San Diego, in a prediction space over the entire continental United States. These accuracies are computed on test sets that were acquired through the same procedures as the training data, so if the acquisition procedures have geographic and demographic biases, then the resulting accuracy estimates will be biased too. Consequently, they may be overly optimistic (or pessimistic!) for some types of authors. In this section, we explore where these text-based geolocation methods are most and least accurate.

5.1 Methods

Our data is drawn from the ten largest metropolitan areas in the United States, and we formulate text-based geolocation as a ten-way classification problem, similar to Han et al. (2014).[6] Using our

[6]Many previous papers have attempted to identify the precise latitude and longitude coordinates of individual authors, but obtaining high accuracy on this task involves much more complex methods, such as latent variable models (Eisenstein et al., 2010; Hong et al., 2012), or multilevel grid structures (Cheng et al., 2010; Roller et al., 2012). Tuning such

user-balanced samples, we apply ten-fold cross validation, and tune the regularization parameter on a development fold, using the vocabulary of the sample as features.

5.2 Results

Many author-attribute prediction tasks become substantially easier as more data is available (Burger et al., 2011), and text-based geolocation is no exception. Since GPS-MSA-BALANCED and LOC-MSA-BALANCED have very different usage rates (Figure 2), perceived differences in accuracy may be purely attributable to the amount of data available per user, rather than to users in one group being inherently harder to classify than another. For this reason, we bin users by the number of messages in our sample of their timeline, and report results separately for each bin. All errorbars represent 95% confidence intervals.

GPS versus location As seen in Figure 6a, there is little difference in accuracy across sampling techniques: the location-based sample is slightly easier to geolocate at each usage bin, but the difference is not statistically significant. However, due to the higher average usage rate in GPS-MSA-BALANCED(see Figure 2), the overall accuracy for a sample of users will appear to be higher on this data.

Demographics Next, we measure classification accuracy by gender and age, using the posterior distribution from the expectation-maximization algorithm to predict the gender of each user (broadly similar results are obtained by using the prior distribution). For this experiment, we focus on the GPS-MSA-BALANCED sample. As shown in Figure 6b, text-based geolocation is consistently more accurate for male authors, across almost the entire spectrum of usage rates. As shown in Figure 6c, older users also tend to be easier to geolocate: at each usage level, the highest accuracy goes to one of the two older groups, and the difference is significant in almost every case. As discussed in § 4, older male users tend to mention many entities, particularly sports-related terms; these terms are apparently more predictive than the non-standard spellings and slang favored by younger authors.

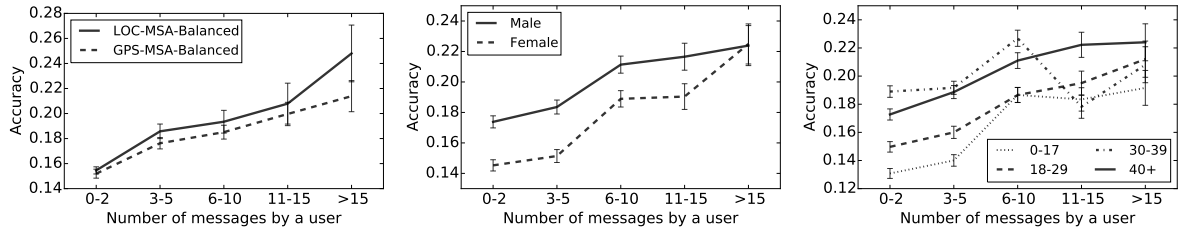
models can be challenging, and the resulting accuracies might be affected by initial conditions or hyperparameters. We therefore focus on classification, employing the familiar and well-understood method of logistic regression.

6 Related Work

Several researchers have studied how adoption of Internet technology varies with factors such as socioeconomic status, age, gender, and living conditions (Zillien and Hargittai, 2009). Hargittai and Litt (2011) use a longitudinal survey methodology to compare the effects of gender, race, and topics of interest on Twitter usage among young adults. Geographic variation in Twitter adoption has been considered both internationally (Kulshrestha et al., 2012) and within the United States, using both the Twitter location field (Mislove et al., 2011) and per-message GPS coordinates (Hecht and Stephens, 2014). Aggregate demographic statistics of Twitter users' geographic census blocks were computed by O'Connor et al. (2010) and Eisenstein et al. (2011b); Malik et al. (2015) use census demographics in spatial error model. These papers draw similar conclusions, showing that the the distribution of geotagged tweets over the US population is not random, and that higher usage is correlated with urban areas, high income, more ethnic minorities, and more young people. However, this prior work did not consider the biases introduced by relying on geotagged messages, nor the consequences for geo-linguistic analysis.

Twitter has often been used to study the geographical distribution of linguistic information, and of particular relevance are Twitter-based studies of regional dialect differences (Eisenstein et al., 2010; Doyle, 2014; Gonçalves and Sánchez, 2014; Eisenstein, 2015) and text-based geolocation (Cheng et al., 2010; Hong et al., 2012; Han et al., 2014). This prior work rarely considers the impact of the demographic confounds, or of the geographical biases mentioned in § 3. Recent research shows that accuracies of core language technology tasks such as part-of-speech tagging are correlated with author demographics such as author age (Hovy and Søgaard, 2015); our results on location prediction are in accord with these findings. Hovy (2015) show that including author demographics can improve text classification, a similar approach might improve text-based geolocation as well.

We address the question about the impact of geographical biases and demographic confounds by measuring differences between three sampling techniques, in both language use and in the accuracy of text-based geolocation. Recent unpublished work proposes reweighting Twitter data to



(a) Classification accuracy by sampling technique

(b) Classification accuracy by user gender

(c) Classification accuracy by imputed age

Figure 6: Classification accuracies

correct biases in political analysis (Choy et al., 2012) and public health (Culotta, 2014). Our results suggest that the linguistic differences between user-supplied profile locations and per-message geotags are more significant, and that accounting for the geographical biases among geotagged messages is not sufficient to offer a representative sample of Twitter users.

7 Discussion

Geotagged Twitter data offers an invaluable resource for studying the interaction of language and geography, and is helping to usher in a new generation of location-aware language technology. This makes critical investigation of the nature of this data source particularly important. This paper uncovers demographic confounds in the linguistic analysis of geo-located Twitter data, but is limited to demographics that can be readily induced from given names. A key task for future work is to quantify the representativeness of geotagged Twitter data with respect to factors such as race and socioeconomic status, while holding geography constant. However, these features may be more difficult to impute from names alone. Another crucial task is to expand this investigation beyond the United States, as the varying patterns of use for social media across countries (Pew Research Center, 2012) implies that the findings here cannot be expected to generalize to every international context.

Acknowledgments Thanks to the anonymous reviewers for their useful and constructive feedback on our submission. The following members of the Georgia Tech Computational Linguistics Laboratory offered feedback throughout the research process: Naman Goyal, Yangfeng Ji, Vinodh Krishan, Ana Smith, Yijie Wang, and Yi Yang. This research was supported by the National Science Foundation under awards IIS-1111142 and RI-1452443, by the National Institutes of Health under award number R01GM112697-01, and by the Air Force Office of Scientific Research. The content is solely the responsibility of the authors and does not necessarily represent the official views of these sponsors.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

David A Broniatowski, Michael J Paul, and Mark Dredze. 2013. National and local influenza surveillance through twitter: An analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Guido Caldarelli, Alessandro Chessa, Fabio Pammolli, Gabriele Pompa, Michelangelo Puliga, Massimo Riccaboni, and Gianni Riotta. 2014. A multi-level geographical study of Italian political elections from Twitter Data. *PloS one*, 9(5):e95809.

Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. ePluribus: Ethnicity on social networks. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 18–25, Menlo Park, California. AAAI Publications.

Zhiyuan Cheng, James Caverlee, and Kyumin Lee. 2010. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 759–768.

Murphy Choy, Michelle Cheong, Ma Nang Laik, and Koo Ping Shung. 2012. Us presidential election 2012 prediction using census corrected twitter model. *arXiv preprint arXiv:1211.0938*.

Aron Culotta. 2014. Reducing sampling bias in social media data for county health inference. In *Joint Statistical Meetings Proceedings*.

Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 98–106, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using Artificial Intelligence*, pages 20–24.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 1277–1287, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011a. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1041–1048, Seattle, WA.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011b. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1365–1374, Portland, OR.

Jacob Eisenstein. 2015. Written dialect variation in online social media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Irene Eleta and Jennifer Golbeck. 2014. Multilingual use of twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.

Nikesh Garera and David Yarowsky. 2009. Modeling latent biographic attributes in conversational genres. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 710–718.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through twitter. *PloS one*, 9(11):e112074.

Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research (JAIR)*, 49:451–500.

Eszter Hargittai and Eden Litt. 2011. The tweet smell of celebrity success: Explaining variation in twitter adoption among a diverse group of young adults. *New Media & Society*, 13(5):824–842.

Brent Hecht and Monica Stephens. 2014. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 197–205, Menlo Park, California. AAAI Publications.

Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of Human Factors in Computing Systems (CHI)*, pages 237–246.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the twitter stream. In *Proceedings of the Conference on World-Wide Web (WWW)*, pages 769–778, Lyon, France.

Dirk Hovy and Anders Søgaard. 2015. Tagging performance correlates with author age. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 483–488, Beijing, China.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 752–762, Beijing, China.

Juhi Kulshrestha, Farshad Kooti, Ashkan Nikravesh, and Krishna P. Gummadi. 2012. Geographic Dissection of the Twitter Network. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, Menlo Park, California. AAAI Publications.

Amanda Lenhart. 2015. Mobile access shifts social media use and other online activities. Technical report, Pew Research Center, April.

P. A. Longley, M. Adnan, and G. Lansley. 2015. The geotemporal demographics of twitter usage. *Environment and Planning A*, 47(2):465–484.

Momin Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*, pages 18–27. The AAAI Press.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*, pages 554–557, Menlo Park, California. AAAI Publications.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Dong Nguyen, Dolf Trieschnigg, A Seza Dogruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1950–1961.

Brendan O'Connor, Jacob Eisenstein, Eric P. Xing, and Noah A. Smith. 2010. A mixture model of demographic lexical variation. In *Proceedings of NIPS Workshop on Machine Learning for Social Computing*, Vancouver.

Pew Research Center. 2012. Social networking popular across globe. Technical report, December.

Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same? characterizing Twitter around the world. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM)*, pages 1025–1030. ACM.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of Workshop on Search and mining user-generated contents*.

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. 2012. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 1500–1510.

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 *Proceedings of the Association for Computational Linguistics (ACL)*, pages 763–772, Portland, OR.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pages 973–981, Honolulu, HI, October. Association for Computational Linguistics.

Svitlana Volkova and Benjamin Van Durme. 2015. Online bayesian models for personal analytics in social media. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 2325–2331.

Benjamin Wing and Jason Baldridge. 2011. Simple supervised document geolocation with geodesic grids. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 955–964, Portland, OR.

Sarita Yardi, Daniel Romero, Grant Schoenebeck, et al. 2009. Detecting spam in a twitter network. *First Monday*, 15(1).

Kathryn Zickuhr. 2013. Location-based services. Technical report, Pew Research Center, Septmeber.

Nicole Zillien and Eszter Hargittai. 2009. Digital distinction: Status-specific types of internet usage*. *Social Science Quarterly*, 90(2):274–291.