Journal of Artificial Intelligence Research 55 (2016) 389-408 Submitted 07/15; published 02/16

Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data

Aron Culotta ACULOTTA@IIT.EDU
Nirmal Kumar Ravi NRAVI@HAWK.IIT.EDU
Department of Computer Science, Illinois Institute of Technology
Chicago, IL 60616

Jennifer Cutler JCUTLER2@STUART.IIT.EDU
Stuart School of Business, Illinois Institute of Technology
Chicago, IL 60616

Abstract

Understanding the demographics of users of online social networks has important applications for health, marketing, and public messaging. Whereas most prior approaches rely on a supervised learning approach, in which individual users are labeled with demographics for training, we instead create a distantly labeled dataset by collecting audience measurement data for 1,500 websites (e.g., 50% of visitors to gizmodo.com are estimated to have a bachelor's degree). We then fit a regression model to predict these demographics from information about the followers of each website on Twitter. Using patterns derived both from textual content and the social network of each user, our final model produces an average held-out correlation of .77 across seven different variables (age, gender, education, ethnicity, income, parental status, and political preference). We then apply this model to classify individual Twitter users by ethnicity, gender, and political preference, finding performance that is surprisingly competitive with a fully supervised approach.

1. Introduction

Social media are increasingly being used to make inferences about the real world, with application to politics (O'Connor, Balasubramanyan, Routledge, & Smith, 2010), health (Dredze, 2012), and marketing (Gopinath, Thomas, & Krishnamurthi, 2014). Understanding the demographic makeup of a sample of social media users is critical to further progress in this area, as it allows researchers to overcome the considerable selection bias in this uncontrolled data. Additionally, this capability will help public messaging campaigns ensure that the target demographic is being reached.

A common approach to demographic inference is supervised classification — from a training set of annotated users, a model is fit to predict user attributes from the content of their writings (Argamon, Dhawle, Koppel, & Pennebaker, 2005; Schler, Koppel, Argamon, & Pennebaker, 2006; Rao, Yarowsky, Shreevats, & Gupta, 2010; Pennacchiotti & Popescu, 2011; Burger, Henderson, Kim, & Zarrella, 2011; Rao, Paul, Fink, Yarowsky, Oates, & Coppersmith, 2011; Al Zamal, Liu, & Ruths, 2012). This approach has a number of limitations: collecting human annotations is costly and error-prone; many demographic variables of interest cannot easily be labeled by inspecting a profile (e.g., income, education level); by restricting learning to a small set of labeled profiles, the generalizability of the classifier is

limited. Additionally, most past work has focused on text as the primary source of evidence, making little use of network evidence.

In this paper, we fit regression models to predict seven demographic variables of Twitter users (age, gender, education, ethnicity, income, parental status, and political preference) based both on whom they follow and on the content of their tweets. Rather than using a standard supervised approach, we construct a distantly labeled dataset consisting of web traffic demographic data from Quantcast.com. By pairing web traffic demographics of a site with the followers of that site on Twitter.com, we fit a regression model between a set of Twitter users and their expected demographic profile. We then evaluate accuracy both in recovering the Quantcast statistics as well as in classifying individual Twitter users. Our experiments investigate several research questions:

RQ1. Can the demographics of a set of Twitter users be inferred from network information alone? We find across seven demographic variables an average held-out correlation of .73 between the web traffic demographics of a website and that predicted by a regression model based on the site's Twitter followers. We can learn, for example, that high-income users are likely to follow The Economist and young users are likely to follow PlayStation.

RQ2. Which is more revealing of demographics: social network features or linguistic features? Overall, we find text-based features to be slightly more predictive than social network features (.79 vs. .73), particularly for income and age variables.

RQ3. Can a regression model be extended to classify individual users? Using a hand-labeled validation set of users annotated with gender, ethnicity, and political preference, we find that a distantly trained regression model provides classification accuracy competitive with a fully-supervised approach. Averaged across the three classification tasks, both approaches obtain an F1 score of 81%.

RQ4. How much follower and linguistic information is needed for prediction? We find that the identities of only 10 friends per user, chosen at random, is sufficient to achieve 90% of the accuracy obtained using 200 friends. Accuracy using linguistic features begins to plateau after 2,000 unique terms are observed per user.

In the remainder of the paper, we first review related work, then describe the data collected from Twitter and QuantCast and the feature representation used for the task; next, we present regression and classification results; finally, we conclude and outline directions for future work.[1]

2. Related Work

Predicting attributes of social media users is a growing area of interest, with recent work focusing on age (Schler et al., 2006; Rosenthal & McKeown, 2011; Nguyen, Smith, & Ros, 2011; Al Zamal et al., 2012), gender (Rao et al., 2010; Burger et al., 2011; Liu & Ruths,

1. Replication code and data are available here: `https://github.com/tapilab/jair-2016-demographics`.

2013), race/ethnicity (Pennacchiotti & Popescu, 2011; Rao et al., 2011), personality (Argamon et al., 2005; Schwartz, Eichstaedt, Kern, Dziurzynski, Ramones, Agrawal, Shah, Kosinski, Stillwell, Seligman, et al., 2013), political affiliation (Conover, Goncalves, Ratkiewicz, Flammini, & Menczer, 2011; Barberá, 2013; Volkova & Van Durme, 2015), and occupation (Preotiuc-Pietro, Lampos, & Aletras, 2015). Other work predicts demographics from web browsing histories (Goel, Hofman, & Sirer, 2012). The majority of these approaches rely on hand-annotated training data, require explicit self-identification by the user, or are limited to very coarse attribute values (e.g., above or below 25-years-old).

Distantly supervised learning (also called lightly or weakly supervised learning) provides an alternative to standard supervised learning — it relies less on individual annotated examples, instead bootstrapping models from declarative constraints. Previous work has developed methods to train classifiers from prior knowledge of label proportions (Jin & Liu, 2005; Musicant, Christensen, & Olson, 2007; Quadrianto, Petterson, & Smola, 2009a; Liang, Jordan, & Klein, 2009; Ganchev, Graca, Gillenwater, & Taskar, 2010; Mann & McCallum, 2010; Zhu, Chen, & Xing, 2014) or prior knowledge of features-label associations (Schapire, Rochery, Rahim, & Gupta, 2002; Druck, Mann, & McCallum, 2008; Melville, Gryc, & Lawrence, 2009). In addition to standard document categorization tasks, lightly supervised approaches have been applied to named-entity recognition (Mann & McCallum, 2010; Ganchev & Das, 2013; Wang & Manning, 2014), dependency parsing (Druck, Mann, & McCallum, 2009; Ganchev, Gillenwater, & Taskar, 2009), language identification (King & Abney, 2013), and sentiment analysis (Melville et al., 2009).

Chang, Rosenn, Backstrom, and Marlow (2010) propose a related distantly supervised approach to demographic inference, inferring user-level ethnicity using name/ethnicity distributions provided by the U.S. Census; however, that approach uses evidence from first and last names, which are often not available, and thus are more appropriate for population-level estimates. Oktay, Firat, and Ertem (2014) extend the work of Chang et al. (2010) to also include statistics over first names. Rao et al. (2011) take a similar approach, also including evidence from other linguistic features to infer gender and ethnicity of Facebook users; they evaluate on the fine-grained ethnicity classes of Nigeria and use very limited training data. More recently, Mohammady and Culotta (2014) trained an ethnicity model for Twitter using county-level supervision, which, like our approach, uses a distant source of supervision to build a model of individual demographics.

There have been several studies predicting population-level statistics from social media. Eisenstein, Smith, and Xing (2011) use geolocated tweets to predict zip-code statistics of race/ethnicity, income, and other variables using Census data; Schwartz et al. (2013) and Culotta (2014) similarly predict county health statistics from Twitter. However, none of this prior work attempts to predict or evaluate at the user level.

The primary methodological novelties of the present work are its use of web traffic data as a form of weak supervision and its use of follower information as the primary source of evidence. Additionally, this work considers a larger set of demographic variables than prior work, and predicts a much more fine-grained set of categories (e.g., six different age brackets instead of two or three used previously).

This paper extends our initial version of this work (Culotta, Kumar, & Cutler, 2015). The novel contributions here include the following: (1) we have added an additional demographic variable (political preference); (2) we have added an additional labeled dataset

for evaluation (also for political preference); (3) whereas the initial version only used friend features, here have introduced textual features from over 9M tweets; (4) we have included a new analysis of how accuracy varies with the number of terms collected per user. We have also reproduced our prior results using newly collected data from QuantCast.com; since QuantCast demographic statistics have changed over time, the overall correlations reported here deviate slightly from those reported in the original version, but the trends and qualitative conclusions remain the same.

3. Data

In this section we describe the various data collected for our experiments.

3.1 Quantcast

Quantcast.com is an audience measurement company that tracks the demographics of visitors to millions of websites. This is accomplished in part by using cookies to track the browsing activity of a large panel of respondents (Kamerer, 2013). As of this writing, the estimated demographics of a large number of websites is publicly accessible through their searchable web interface.

We sampled 1,532 websites from Quantcast and downloaded statistics for seven demographic variables:

- **Gender:** Male, Female
- **Age:** 18-24, 25-34, 35-44, 45-54, 55-64, 65+
- **Income:** $0-50k, $50-100k, $100-150k, $150k+
- **Education:** No College, College, Grad School
- **Children:** Kids, No Kids
- **Ethnicity:** Caucasian, Hispanic, African American, Asian
- **Political Preference:** Democrat, Republican

For each variable, Quantcast reports the estimated percentage of visitors to a website with a given demographic.

3.2 Twitter

For each website collected in the previous step, we executed a script to search for its Twitter account, then manually verified it; 1,066 accounts from the original set of 1,532 were found. An assumption of this work is that the demographic profiles of followers of a website on Twitter are correlated with the demographic profiles of visitors to that website. While there are undoubtedly biases introduced here (e.g., Twitter users may skew younger than the web traffic panel), in aggregate these differences should have limited impact on the final model.

We represent each of the 1,066 Twitter accounts with feature vectors derived from information about their followers. Below, we describe features based both on the social network and on the linguistic content of each follower's tweets.

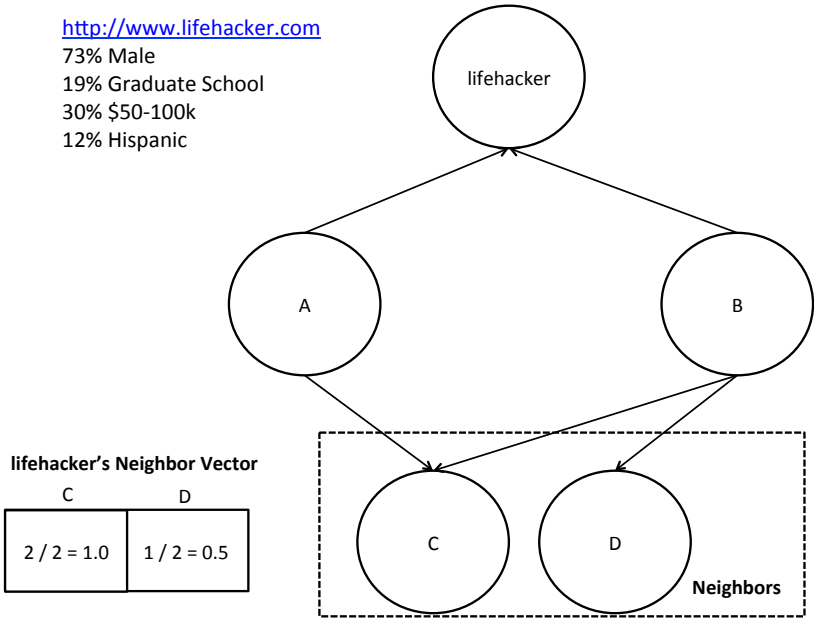

Figure 1: Data model. We collect QuantCast demographic data for each website, then construct a **Neighbor Vector** from the Twitter connections of that website, based on the proportion of the website's followers that are friends with each neighbor.

3.2.1 Friend Features

Recall that Twitter users are allowed to "follow" other accounts, which introduces an asymmetric relationship between users. If X follows Y, we say that Y is a *friend* of X (though the reverse may not be true). For each account, we queried the Twitter REST API to sample 300 of its followers, using the `followers/ids` request. This sample is not necessarily uniform. The Twitter API documentation states that "At this time, results are ordered with the most recent following first — however, this ordering is subject to unannounced change and eventual consistency issues."

For each of these followers, we then collected up to 5,000 of the accounts they follow, called *friends*, using the `friends/ids` API request. Thus, for each of the original accounts from Quantcast, we have up to $(300 * 5K = 1.5M)$ additional accounts that are two hops from the original account (the friend of a follower). We refer to these discovered accounts as *neighbors* of the original Quantcast account. Of course, many of these accounts will be duplicates, as two different followers will follow many of the same accounts (i.e., *triadic closure*) — indeed, our core assumption is that the number of such duplicates represents the strength of the similarity between the neighbors.

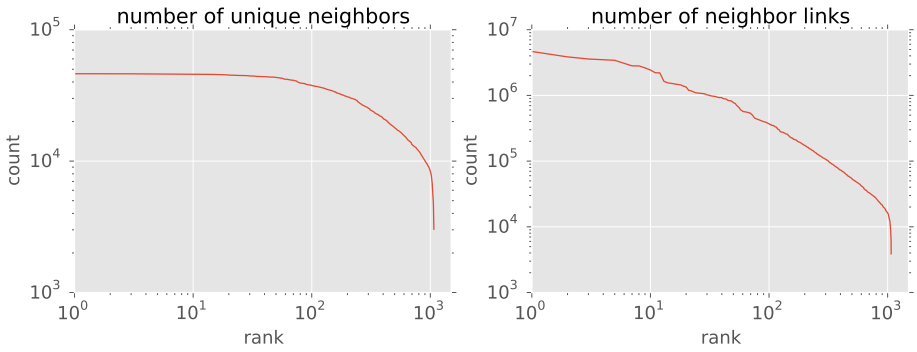

Figure 2: Rank-order frequency plots of the number of neighbors per account and the number of links to all neighbors per account.

For each of the original accounts, we compute the fraction of its followers that are friends with each of its neighbors and store this in a *neighbor vector*. Figure 1 shows an example. Suppose a Quantcast account LifeHacker has two followers A and B; and that A follows C, and B follows C and D. Then the neighbor vector for LifeHacker is $\{(C, 1), (D, .5)\}$. This suggests that LifeHacker has a stronger relationship with C than D.[2] For example, in our data the top neighbors of LifeHacker are EA (the video game developer), GTASeries (a video game fan site), and PlayStation.

The resulting dataset consists of 1.7M unique neighbors of the original 1,066 accounts. To reduce dimensionality, we removed neighbors with fewer than 100 followers, leaving 46,649 unique neighbors with a total of 178M incoming links. Figure 2 plots the number of unique neighbors per account as well as the number of neighbor links per account.

3.2.2 Text Features

In addition to the neighbor vector, we created an analogous vector based on the tweets of the followers of each account. We collected the most recent 200 tweets for each of the 300 followers of each of the 1,066 accounts using the `statuses/user_timeline` API request.

For each tweet, we perform standard tokenization, removing non-internal punctuation, converting to lower case, and maintaining hashtag and mentions. URLs are collapsed to a single feature type, as are digits (e.g., "12" is mapped to "99"; "123" is mapped to "999"). Characters repeated more than twice are converted to a single occurrence. Terms used by fewer than 20 different users are removed.

The resulting dataset consists of 9,427,489 tweets containing 112,642 unique terms written by 59,431 users. For each of the original 1,066 accounts, we create a *text vector* similar to the previous section. Each value represents the proportion of followers of the account who use that term. E.g., $x_{ij} = .1$ indicates that 10% of the 300 followers of account i use term j.

2. Note that we use friends of A rather than followers, since friend links are created by A and are thus more likely to indicate A's interests.

4. Analysis

In this section, we report results predicting demographics both at the aggregate level (i.e., the demographics of an account's followers) and at the user level (i.e., an individual Twitter user's demographic profile).

4.1 Regression

For each Quantcast site, we pair its demographic variables with its friend and text feature vectors to construct a regression problem. Thus, we attempt to predict the demographic profile of the followers of a Twitter account based on the friends of those followers and the content of their tweets.

Due to the high dimensionality (46,649 friend features and 112,642 text features) and small number of examples (1,066), we use elastic net regularization (Zou & Hastie, 2005), which combines both L1 and L2 penalties. Furthermore, since each output variable consists of dependent categories (e.g., age brackets), we use a multi-task variant of elastic net to ensure that the same features are selected by the L1 regularizer for each category. We use the implementation of `MultiTaskElasticNet` in `scikit-learn` (Pedregosa et al., 2011).

Recall that standard linear regression selects coefficients β to minimize the squared error on a list of training instances $\{\mathbf{x}_i, y_i\}_{i=1}^N$, for feature vector $\mathbf{x}_i$ and expected output y_i.

$$\beta^* \leftarrow \text{argmin}_\beta \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T \mathbf{x}_i)^2$$

Lasso imposes an L1 regularizer on β, while ridge regression imposes an L2 regularizer on β. Elastic net combines both penalties:

$$\beta^* \leftarrow \text{argmin}_\beta \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta^T \mathbf{x}_i)^2 + \lambda_1 ||\beta||_1 + \lambda_2 ||\beta||_2^2$$

where λ_1 and λ_2 control the strength of L1 and L2 regularizers, respectively. The L1 regularizer encourages sparsity (i.e., many 0 values in β), while the L2 regularizer prevents β values from becoming too large.

Multi-task elastic net extends standard elastic net to groups of related regression problems (Obozinski & Taskar, 2006). E.g., in our case, we would like to account for the fact that the regressions for "No College", "College", and "Grad School" are related; thus, we would like the sparse solutions to be similar across tasks (that is, L1 should select the same features for each task).

Let $\beta^{(j)}$ be the coefficients for task j, and let $\beta_k = (\beta_k^{(1)} \dots \beta_k^{(M)})^T$ be the vector of coefficients formed by concatenating the coefficients for the kth feature across all M tasks. Then multi-task elastic net objective enforces that similar features are selected across tasks:

$$\beta^* \leftarrow \text{argmin}_\beta \sum_{j=1}^{M} \frac{1}{N_j} \sum_{i=1}^{N_j} (y_i^{(j)} - \beta^{(j)T} \mathbf{x}_i^{(j)})^2 + \lambda_1 \sum_{k=1}^{p} ||\beta_k||_1 + \lambda_2 ||\beta||_2^2$$

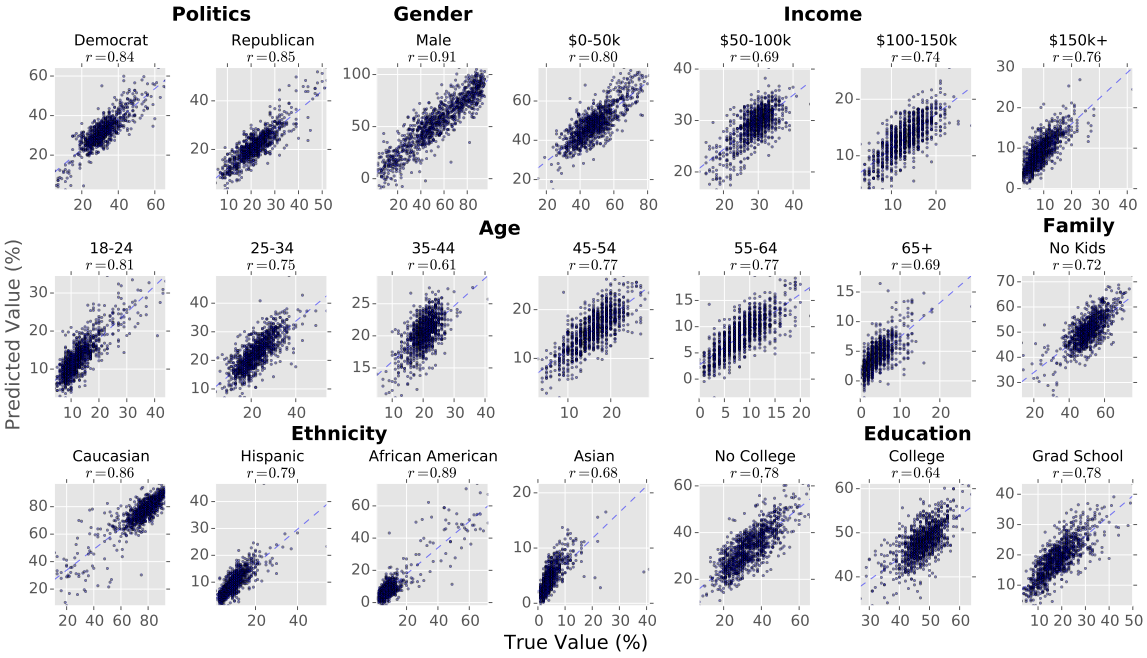

Figure 3: Scatter plots of the true demographic variables from Quantcast versus those predicted using elastic net regression fit to friend and text features from Twitter. The predictions are computed using five fold cross-validation; each panel also reports the held-out correlation coefficient (r).

where N_j is the number of instances for task j and p is the number of features.

We fit three versions of this model using three different feature sets: Friends, Text, and Friends+Text. We tuned the regularization parameters on a held-out set of 200 accounts for Gender prediction, setting the `scikit-learn` parameters l1_ratio=0.5 for each model, alpha=$1e{-}5$ for the Friends model, and alpha=$1e{-}2$ for the Text and Friends+Text models.

4.1.1 Regression Results

We perform five-fold cross-validation and report the held-out correlation coefficient (r) between the predicted and true demographic variables. Figure 3 displays the resulting scatter plots for each of the 21 categories for 7 demographic variables.

We can see that overall the correlation is very strong: .77 on average, ranging from .61 for the 35-44 age bracket to .91 for Male. All of these correlation coefficients are significant using a two-tailed t-test ($p < 0.01$), with a Bonferroni adjustment for the 21 comparisons. These results indicate that the neighbor and text vectors provides a reliable signal of the demographics of a group of Twitter users. To put this correlation value in context, as an indirect comparison, Eisenstein et al. (2011) predict the ethnicity proportions by ZIP Code using Twitter data and obtain a maximum correlation of .337 (though the data and experimental setup differ considerably).

Category	Value	Top Accounts
Gender	Male	AdamSchefter, SportsCenter, espn, mortreport, WIRED
	Female	TheEllenShow, Oprah, MarthaStewart, Pinterest, Etsy
Age	18-24	IGN, PlayStation, RockstarGames, Ubisoft, steam_games
	25-34	azizansari, lenadunham, mindykaling, WIRED
	35-44	cnnbrk, BarackObama, AP, TMZ, espn
	45-54	FoxNews, cnnbrk, AP, WSJ, CNN
	55-64	FoxNews, cnnbrk, AP, ABC, WSJ
	65+	FoxNews, AP, WSJ, cnnbrk, DRUDGE_REPORT
Income	$0-50k	YouTube, PlayStation, IGN, RockstarGames, Drake
	$50-100k	AdamSchefter, cnnbrk, SportsCenter, espn, ErinAndrews
	$100-150k	WSJ, espn, AdamSchefter, SportsCenter, ErinAndrews
	$150k+	WSJ, TheEconomist, Forbes, nytimes, business
Politics	Democrat	BarackObama, Oprah, NewYorker, UncleRUSH, MichelleObama
	Republican	FoxNews, michellemalkin, seanhannity, megynkelly, DRUDGE_REPORT
Education	No College	YouTube, PlayStation, RockstarGames, Xbox, IGN
	College	StephenAtHome, WIRED, ConanOBrien, mashable
	Grad School	nytimes, WSJ, NewYorker, TheEconomist, washingtonpost
Children	No Kids	NewYorker, StephenAtHome, nytimes, maddow, pitchfork
	Has Kids	parenting, parentsmagazine, HuffPostParents, TheEllenShow, thepioneerwoman
Ethnicity	Caucasian	jimmyfallon, FoxNews, blakeshelton, TheEllenShow, TheOnion
	Hispanic	latimes, Lakers, ABC7, Dodgers, KTLA
	Afr. Amer.	KevinHart4real, Drake, Tip, iamdiddy, UncleRUSH
	Asian	TechCrunch, WIRED, BillGates, TheEconomist, SFGate

Table 1: Accounts with the highest estimated coefficients for each category.

To further examine these results, Table 1 displays the features with the 5 largest coefficients per class according to the regression model fit using only friend features. Many results match common stereotypes: sports accounts are predictive of men (AdamShefter and MortReport are ESPN reporters), video game accounts are predictive of younger users (IGN is a video gaming media company), financial news accounts are predictive of greater income, and parenting magazines are predictive of users who have children. There also appear to be some geographic effects, as California-related accounts are highly weighted for both Hispanic and Asian categories. There seems to be good city-level resolution — Los Angeles accounts (*latimes*, *Lakers*) are more strongly correlated with Hispanic users, whereas San Francisco accounts (*SFGate*, *SFist*, *SFWeekly*) are more strongly correlated with Asian users. There does seem to be some selection bias, so one must use caution in interpreting the results. For example, BillGates is predictive of Asian users, but this is in part because California has many Asian-Americans and in part because California has a very strong technology sector.

Category	Value	Top Terms
Gender	Male	film, guy, gay, man, fuck, game, team, internet, review, guys
	Female	hair, her, omg, family, girl, she, girls, cute, beautiful, thinking
Age	18-24	d, haha, album, x, xd, _:, actually, stream, wanna, im
	25-34	super, dc, baby, definitely, nba, pregnancy, wedding, even, entire, nyc
	35-44	star, fans, kids, tv, bike, mind, store, awesome, screen, son
	45-54	wow, vote, american, comes, ca, santa, county, boys, nice, high
	55-64	vote, golf, red, american, country, north, county, holiday, smile, 99,999
	65+	vote, golf, @foxnews, holiday, may, american, he, family, north, national
Income	$0-50k	lol, games, @youtube, damn, black, ps9, side, d, community, god
	$50-100k	great, seattle, he, performance, lose, usa, kansas, iphone, wow, cold
	$100-150k	santa, flight, nice, looks, practice, congrats, bike, dc, retweet, ride
	$150k+	dc, nyc, market, @wsj, congrats, beach, san, york, ca, looks
Politics	Democrat	women, u, ain't, nyc, equality, la, voice, seattle, dc, @nytimes
	Republican	@foxnews, christmas, #tcot, football, county, morning, family, christians, country, obama's
Education	No College	lol, games, put, @youtube, county, made, ps9, xbox, videos, found
	College	our, you're, seattle, photo, @mashable, la, apple, fashion, probably, san
	Grad School	dc, @nytimes, market, which, review, excellent, boston, also, congrats, @washingtonpost
Children	No Kids	care, street, gay, years, health, drink, dc, white, ht, album
	Has Kids	kids, school, child, family, kid, daughter, children, utah, moms, parents
Ethnicity	Caucasian	christmas, fun, dog, country, st, could, luck, guy, florida, john
	Hispanic	la, los, san, el, angeles, california, ca, lol, l.a, lakers
	Afr. Amer.	black, lol, bout, ain't, brown, lil, african, blessed, smh, atlanta
	Asian	chinese, la, sf, san, china, korea, india, bay, vs, hi

Table 2: Terms with the highest estimated coefficients for each category.

Additionally, Table 2 shows the top 10 terms for each category from the *text-only* model. These text features follow many of the same trends as the friend features, perhaps with a greater level of granularity. While most terms are self-evident, we highlight a few here: "xd" is an emoticon for laughing used among young users, and is often separated by spaces (thus the tokens "x" and "d" appearing separately in the 18-24 bracket); "smh" stands for

Model	Friends	Text	Friends + Text	Average
multi-task elastic net	**.73**	**.79**	.77	**.76**
elastic net	.72	**.78**	.76	.75
ridge	.62	**.79**	**.78**	.73

Table 3: Average held-out correlation across all demographic variables for three competing regression models.

"shaking my head", an expression of disbelief that is predictive of African American users; "hi" is an abbreviation for the state of Hawaii, which has a large Asian population.

Finally, we compare multi-task elastic net with the single-task variant of elastic net and ridge regression (with regularization parameters tuned as before). Table 3 shows that the three methods mostly produce comparable accuracy, with the exception of the friends features, in which ridge regression performs substantially worse than the others.

4.2 Classification

The regression results suggest that the proposed model can accurately characterize the demographics of a group of Twitter accounts. In this section, we provide additional validation with manually annotated Twitter accounts to investigate whether the same model can accurately predict the demographics of individual users.

4.2.1 Labeled Data

Many of the demographic variables are difficult to label at the individual level — e.g., income or education level is rarely explicitly mentioned in either a profile or tweet. Indeed, an advantage of the approach here is that aggregate statistics are more readily available for many demographics of interest that are difficult to label at the individual level. For validation purposes, we focus on three variables that can fairly reliably be labeled for individuals: gender, ethnicity, and political preference.

The gender and ethnicity data were originally collected by Mohammady and Culotta (2014) as follows: First, we used the Twitter Streaming API to obtain a random sample of users, filtered to the United States (using time zone and the place country code from the profile). From six days' worth of data (December 6-12, 2013), we sampled 1,000 profiles at random and categorized them by analyzing the profile, tweets, and profile image for each user. We categorized 770 Twitter profiles into one of four ethnicities (Asian, African American, Hispanic, Caucasian). Those for which ethnicity could not be determined were discarded (230/1,000; 23%).[3] The category frequency is Asian (22), African American (263), Hispanic (158), Caucasian (327). To estimate inter-annotator agreement, a second annotator sampled and categorized 120 users. Among users for which both annotators selected one of the four categories, 74/76 labels agreed (97%). There was some disagreement over when the category could be determined: for 21/120 labels (17.5%), one annotator indicated the category could not be determined, while the other selected a category. Gender

3. This introduces some bias towards accounts with identifiable ethnicity; we leave an investigation of this for future work.

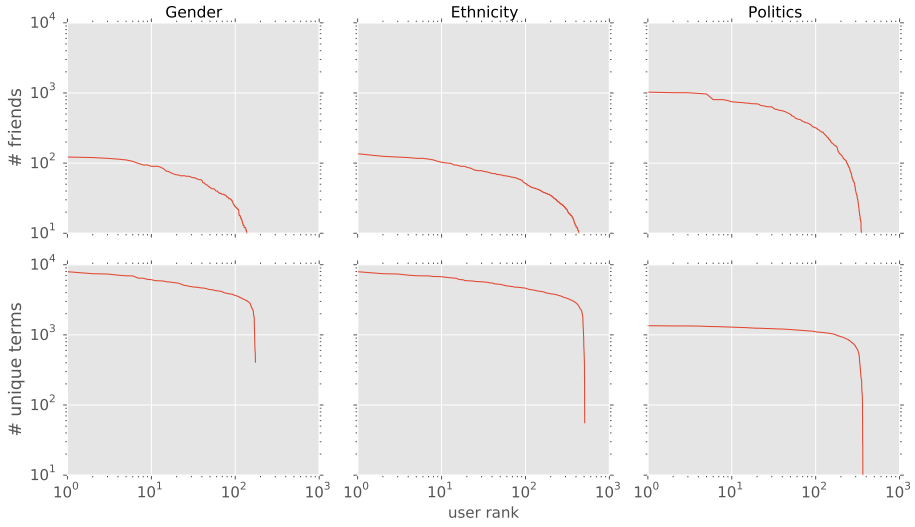

Figure 4: Rank-order frequency plots of the number of friends per user and the number of unique terms per user in each of the labeled datasets (gender, ethnicity, politics). These friends and terms are restricted to one of the 46,649 accounts and 112,642 terms used in the regression experiments.

annotation was done automatically by comparing the first name provided in the user profile with the U.S. Census list of names by gender (Census, 1990). Ambiguous names were removed.

For each user, we collected up to 200 of their friends using the Twitter API. We removed accounts that restricted access to friend information; we also removed the Asian users due to the small sample size, leaving a total of 615 users. For classification, each user is represented by the identity of their friends (up to 200). Only those friend accounts contained in the 46,649 accounts used for the regression experiments were retained. We additionally collected up to 3,200 tweets from each user and constructed a binary term vector, using the same tokenization as in the regression model.

The political preference data comes from Volkova, Coppersmith, and Van Durme (2014), who in turn builds on the labeled data of Pennacchiotti and Popescu (2011) and Al Zamal et al. (2012). Volkova (2014) provides a detailed description of the data. We use the *geo-centric* portion of the data, which contains Twitter users from Maryland, Virginia, or Delaware who report their political affiliation in their Twitter profile description (e.g., "I'm a father, husband, Republican"). Note that our feature representation does not consider tokens from the user profile. This contains 183 Republican users and 230 Democratic users. (We do not consider related political datasets that were annotated based on whom the user follows, as this may give an unfair advantage to the friend features.) Each user has up to 5,000 friends and 200 tweets. Figure 4 shows the number of friends and the number of unique terms per user for each dataset.

	Friends		Text		Friends + Text	
	distant	**full**	**distant**	**full**	**distant**	**full**
Gender	.75	.66	.86	.84	**.87**	.84
Ethnicity	.60	.68	**.86**	**.86**	.81	**.86**
Politics	.80	**.83**	.56	.73	.74	.73
Average	.72	.72	.76	**.81**	**.81**	**.81**

Table 4: F1 results for Twitter user classification on manually annotated data. We consider three different feature sets (Friends, Text, Friends+Text), as well as two classification models: `full` is a fully-supervised logistic regression classifier fit to manually labeled Twitter users; `distant` is our proposed distantly-supervised regression model, fit only on QuantCast data, using no manually annotated Twitter users. The largest values in each row are in bold.

4.2.2 Classification Models

As our model was initially trained for a regression task, we make a few modifications to apply it to a classification task. We represent each user in the labeled data as a binary vector of friend and text features, using the same tokenization as in the regression results. For example, if a user follows accounts A and B, then the feature values are 1 for those corresponding accounts; similarly, if the user mentions terms X and Y, then those feature values are 1. To repurpose the regression model to perform classification, we must modify the coefficients returned by regression. We first compute the z-score of each coefficient with respect to the other coefficients for that category value. E.g., all coefficients for the *Male* class are adjusted to have mean 0 and unit variance. This makes the coefficients comparable across labels. To classify each user, we then compute the dot product between the coefficients and the binary feature vector, selecting the class with maximum value.

The regression model fit on the combined Friend+Text feature set performed poorly in initial classification experiments. Upon investigation, we determined that the coefficients for the two types of feature tended to differ by an order of magnitude. Rather than use this model directly, we instead adopted an ensemble approach by combining the outputs of the two models trained separately on text and friend features. To classify a set of users, we computed the feature-coefficient dot product separately for the text and friend models, then computed the z-score of the resulting values by class label (e.g., all dot-products produced by the text model for the Male class were standardized to have zero mean and unit variance). This put the predicted values for each model in the same range. We finally summed the outputs of both models and returned the class with the maximum value for each user.

We also compared with a fully-supervised baseline. We trained a logistic regression classifier with L2 regularization, using the same feature representation as above. We perform three-fold cross-validation to compare accuracy with the distantly supervised approach.

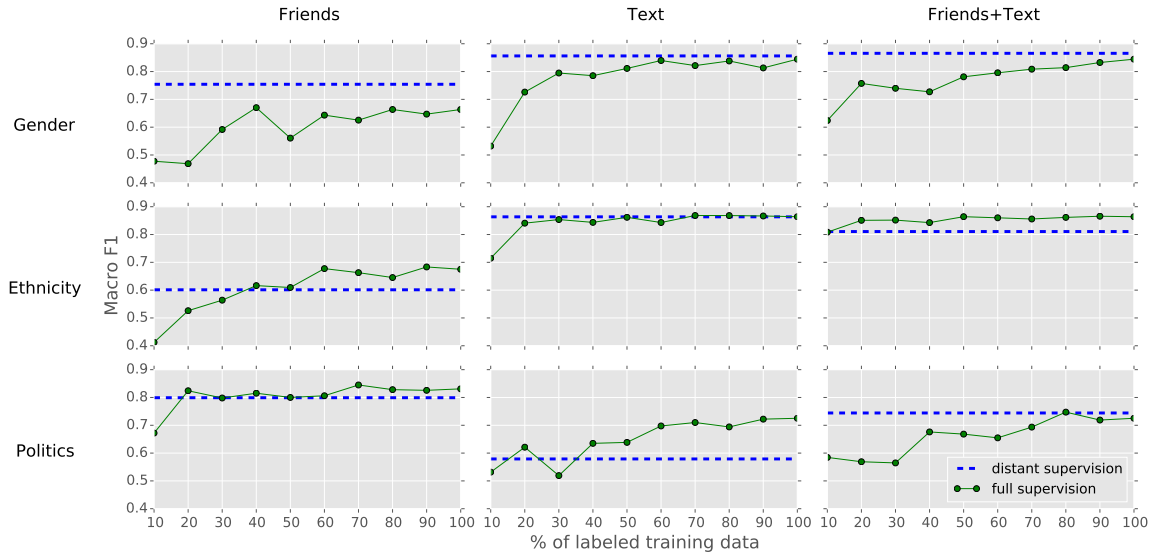

Figure 5: Twitter user classification results comparing a standard logistic regression classifier (`full supervision`), trained using cross-validation, versus the proposed approach (`distant supervision`), which is fit solely on statistics from Quantcast, with no individually labeled data. Distant supervision is comparable to full supervision, even after 100% of the available training data are provided.

4.2.3 Classification Results

Table 4 compares the F1 scores for our distantly supervised approach (`distant`) as well as the fully supervised baseline (`full`). We report results using each of the three feature sets for the three labeled datasets.

Overall, the distantly supervised approach is comparable to the fully supervised approach. For two of the three tasks, the F1 score of `distant` meets or exceeds `full`. For the third task (politics), the best `distant` method is within 3% of the best `full` method (.80 vs .83). Averaging over all three tasks, the best `distant` method is indistinguishable from the best `full` method (both produce F1 scores of 81%).

The primary result in which distant supervision performs poorly is political classification using text features. We speculate that this is in part due to the small number of tweets per user available in this dataset (at most 200 tweets per user). Additionally, the text features used in the distantly supervised approach were collected several years after the tweets contained in the political dataset. Given the rapid topical changes of political dialogue, it is likely that many of the highly-weighted terms in the `distant` text model are less relevant to this older data. The results do suggest that friend features may be less susceptible to such data drift – the friend-based `distant` model performs much better than the text-based model (.80 vs .56).

As the number of labeled data are relatively small (fewer than 1,000 users), we examined the accuracy of the fully supervised approach as the number of labeled data increase (Figure 5). It appears that supervised classification accuracy has mostly plateaued on each

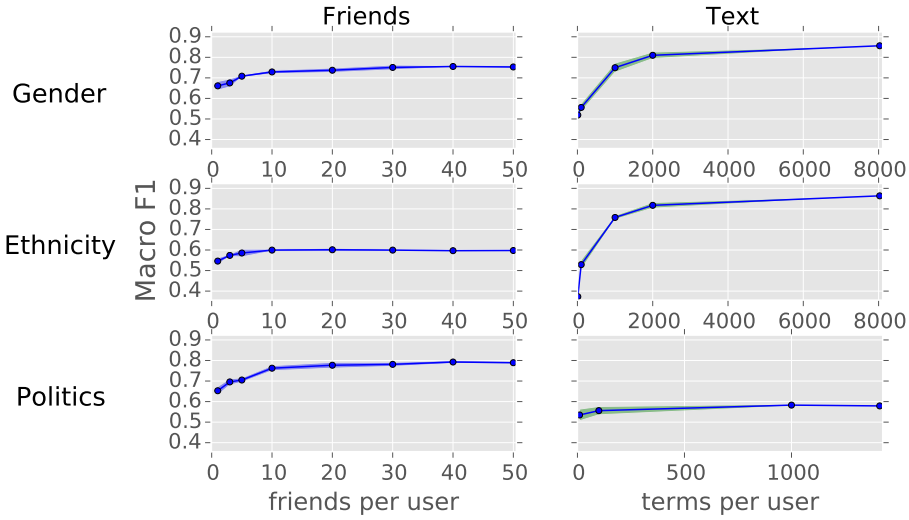

Figure 6: Classification F1 scores of the distantly supervised approach as the number of friends and number of unique terms collected per user increase (with standard deviations computed over five random trials).

task (with a possible exception of the Friends+Text features for gender classification). For ethnicity, `distant` outperforms `full` until over half of the labeled data is used to fit the classification approach, after which `full` dominates. Thus, it appears that the distantly supervised approach is comparable to fully supervised learning across a range of sizes of labeled data.

4.2.4 Sensitivity to Number of Features

Finally, we investigate how much information we need about a user before we can make an accurate prediction of their demographics. To do so, we perform an experiment in which we randomly sample a subset of friends and terms for each user, and we report F1 as the number of selected features increases. For friends, we consider subsets of size $\{1, 3, 5, 10, 20, 30, 40, 50\}$ (values greater than 50 did not significantly increase accuracy). For terms, we consider subsets of size $\{10, 100, 1000, 2000, 8029\}$ (8,029 is the maximum number of unique terms used by any single user in the labeled data).

Figure 6 displays the results. We can see that accuracy plateaus quickly using friend features: for all three tasks, the F1 score using only 10 friends is within 5% of the score using all 200 friends. For text features, accuracy begins to plateau at around 2K unique terms for the Gender and Ethnicity tasks. The lower accuracy using Text features for Politics is likely due in part to the simple fact that the Politics data have fewer tweets per user (a maximum of 200 tweets per user, compared to up to 3,200 for the Gender and Ethnicity tasks).

These results have implications for scalability — Twitter API rate limits make it difficult to collect the complete social graph or tweets for a set of users. Additionally, this has

important privacy implications; revealing even a small amount of social information may also reveal a considerable amount of demographic information. Twitter users concerned about privacy may wish to disable the setting that makes friend identity information public.

5. Conclusions and Future Work

In this paper, we have shown that pairing web traffic demographic data with Twitter data provides a simple and effective way to train a demographic inference model without any annotation of individual profiles. We have validated the approach both in aggregate (by comparing with Quantcast data) and at the individual level (by comparing with hand-labeled annotations), finding high accuracy in both cases. Somewhat surprisingly, the approach outperforms a fully-supervised approach for gender classification, and is competitive for ethnicity and political classification.

In short-term future work, we will test the generalizability of this approach to new groups of Twitter users. For example, we can collect users by city or county and compare the predictions with the Census demographics from that geographic location. Additionally, we will investigate ways to combine labeled and unlabeled data using semi-supervised learning (Quadrianto, Smola, Caetano, & Le, 2009b; Ganchev et al., 2010; Mann & McCallum, 2010). Finally, to fully validate across all demographic variables, we will consider administering surveys to Twitter users to compare predictions with self-reported survey responses.

Additional future work may investigate more sophisticated types of distant supervision. For example, homophily constraints can be imposed to encourage neighbors to have similar demographics; location constraints can use be used to learn from county demographic data. Also, while the multi-task model captures some interaction between demographic variables at training time, we can also use collective inference to reflect the correlations among demographic variables. Finally, we have only considered a simple bag-of-words feature representations; future work may investigate low-dimensional embeddings and non-linear models.

Acknowledgments

This research was funded in part by support from the IIT Educational and Research Initiative Fund. Culotta was supported in part by the National Science Foundation under grant #IIS-1526674. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

References

Al Zamal, F., Liu, W., & Ruths, D. (2012). Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Argamon, S., Dhawle, S., Koppel, M., & Pennebaker, J. W. (2005). Lexical predictors of personality type. In *In proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Barberá, P. (2013). Birds of the same feather tweet together. bayesian ideal point estimation using twitter data. In *Proceedings of the Social Media and Political Participation, Florence, Italy*, pp. 10–11.

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 1301–1309, Stroudsburg, PA, USA. Association for Computational Linguistics.

Census (1990). List of surnames. `http://www2.census.gov/topics/genealogy/1990surnames`. Accessed: 2015-06-01.

Chang, J., Rosenn, I., Backstrom, L., & Marlow, C. (2010). ePluribus: ethnicity on social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). Predicting the political alignment of twitter users. In *IEEE Third international conference on social computing (SOCIALCOM)*, pp. 192–199. IEEE.

Culotta, A. (2014). Estimating county health statistics with twitter. In *CHI*.

Culotta, A., Kumar, N. R., & Cutler, J. (2015). Predicting the demographics of twitter users from website traffic data. In *Twenty-ninth National Conference on Artificial Intelligence (AAAI)*.

Dredze, M. (2012). How social media will change public health. *IEEE Intelligent Systems*, *27*(4), 81–84.

Druck, G., Mann, G., & McCallum, A. (2008). Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 595–602.

Druck, G., Mann, G., & McCallum, A. (2009). Semi-supervised learning of dependency parsers using generalized expectation criteria. In *ACL*.

Eisenstein, J., Smith, N. A., & Xing, E. P. (2011). Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 1365–1374, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ganchev, K., & Das, D. (2013). Cross-lingual discriminative learning of sequence models with posterior regularization.. In *EMNLP*, pp. 1996–2006.

Ganchev, K., Gillenwater, J., & Taskar, B. (2009). Dependency grammar induction via bitext projection constraints. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pp. 369–377. Association for Computational Linguistics.

Ganchev, K., Graca, J., Gillenwater, J., & Taskar, B. (2010). Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, *11*, 2001–2049.

Goel, S., Hofman, J. M., & Sirer, M. I. (2012). Who does what on the web: A large-scale study of browsing behavior.. In *ICWSM*.

Gopinath, S., Thomas, J. S., & Krishnamurthi, L. (2014). Investigating the relationship between the content of online word of mouth, advertising, and brand performance. *Marketing Science*, *33*(2), 241–258.

Jin, R., & Liu, Y. (2005). A framework for incorporating class priors into discriminative classification. In *In PAKDD*.

Kamerer, D. (2013). Estimating online audiences: Understanding the limitations of competitive intelligence services. *First Monday*, *18*(5).

King, B., & Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*, pp. 1110–1119.

Liang, P., Jordan, M. I., & Klein, D. (2009). Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, p. 641648, New York, NY, USA. ACM.

Liu, W., & Ruths, D. (2013). What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium on Analyzing Microtext*.

Mann, G. S., & McCallum, A. (2010). Generalized expectation criteria for semi-supervised learning with weakly labeled data. *J. Mach. Learn. Res.*, *11*, 955–984.

Melville, P., Gryc, W., & Lawrence, R. D. (2009). Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, p. 12751284, New York, NY, USA. ACM.

Mohammady, E., & Culotta, A. (2014). Using county demographics to infer attributes of twitter users. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*.

Musicant, D., Christensen, J., & Olson, J. (2007). Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007*, pp. 252–261.

Nguyen, D., Smith, N. A., & Ros, C. P. (2011). Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Scie nces, and Humanities*, LaTeCH '11, pp. 115–123, Stroudsburg, PA, USA. Association for Computational Linguistics.

Obozinski, G., & Taskar, B. (2006). Multi-task feature selection. In *In the workshop of structural Knowledge Transfer for Machine Learning in the 23rd International Conference on Machine Learning (ICML*.

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

Oktay, H., Firat, A., & Ertem, Z. (2014). Demographic breakdown of twitter users: An analysis based on names. In *Academy of Science and Engineering (ASE)*.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Machine Learning Research*, *12*, 2825–2830.

Pennacchiotti, M., & Popescu, A.-M. (2011). A machine learning approach to twitter user classification.. In Adamic, L. A., Baeza-Yates, R. A., & Counts, S. (Eds.), *ICWSM*. The AAAI Press.

Preotiuc-Pietro, D., Lampos, V., & Aletras, N. (2015). An analysis of the user occupational class through twitter content. In *ACL*.

Quadrianto, N., Petterson, J., & Smola, A. J. (2009a). Distribution matching for transduction. In *Advances in Neural Information Processing Systems 22*, p. 15001508. MIT Press.

Quadrianto, N., Smola, A. J., Caetano, T. S., & Le, Q. V. (2009b). Estimating labels from label proportions. *J. Mach. Learn. Res.*, *10*, 2349–2374.

Rao, D., Paul, M. J., Fink, C., Yarowsky, D., Oates, T., & Coppersmith, G. (2011). Hierarchical bayesian models for latent attribute detection in social media. In *ICWSM*.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, pp. 37–44, New York, NY, USA. ACM.

Rosenthal, S., & McKeown, K. (2011). Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pp. 763–772, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schapire, R. E., Rochery, M., Rahim, M. G., & Gupta, N. K. (2002). Incorporating prior knowledge into boosting. In *Proceedings of the Nineteenth International Conference*, pp. 538–545.

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 06–03.

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. (2013). Characterizing geographic variation in well-being using tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Volkova, S. (2014). Twitter data collection: Crawling users, neighbors and their communication for personal attribute prediction in social media. Tech. rep., Johns Hopkins University.

Volkova, S., Coppersmith, G., & Van Durme, B. (2014). Inferring user political preferences from streaming communications. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Volkova, S., & Van Durme, B. (2015). Online bayesian models for personal analytics in social media. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, Austin, TX.

Wang, M., & Manning, C. D. (2014). Cross-lingual projected expectation regularization for weakly supervised learning. *TACL*, *2*, 55–66.

Zhu, J., Chen, N., & Xing, E. P. (2014). Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, *15*, 1799–1847.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.