

# Mining the Demographics of Political Sentiment from Twitter Using Learning from Label Proportions

Ehsan Mohammady Ardehaly  
Department of Computer Science  
Illinois Institute of Technology  
Chicago, Illinois 60616  
Email: emohamm1@hawk.iit.edu

Aron Culotta  
Department of Computer Science  
Illinois Institute of Technology  
Chicago, Illinois 60616  
Email: aculotta@iit.edu

*Abstract*—Opinion mining and demographic attribute inference have many applications in social science. In this paper, we propose models to infer daily joint probabilities of multiple latent attributes from Twitter data, such as political sentiment and demographic attributes. Since it is costly and time-consuming to annotate data for traditional supervised classification, we instead propose scalable Learning from Label Proportions (LLP) models for demographic and opinion inference using U.S. Census, national and state political polls, and Cook partisan voting index as population level data. In LLP classification settings, the training data is divided into a set of unlabeled bags, where only the label distribution in of each bag is known, removing the requirement of instance-level annotations. Our proposed LLP model, Weighted Label Regularization (WLR), provides a scalable generalization of prior work on label regularization to support weights for samples inside bags, which is applicable in this setting where bags are arranged hierarchically (e.g., county-level bags are nested inside of state-level bags). We apply our model to Twitter data collected in the year leading up to the 2016 U.S. presidential election, producing estimates of the relationships among political sentiment and demographics over time and place. We find that our approach closely tracks traditional polling data stratified by demographic category, resulting in error reductions of 28-44% over baseline approaches. We also provide descriptive evaluations showing how the model may be used to estimate interactions among many variables and to identify linguistic temporal variation, capabilities which are typically not feasible using traditional polling methods.

## I. INTRODUCTION

Recent research has demonstrated the feasibility of estimating quantities of public interest from online social network data, with applications to health [1], politics [2] and marketing [3]. However, practitioners are often more interested in investigating the interactions among sets of variables, rather than estimating the trend of a single variable. For example, health researchers may want to know not only what the influenza rate is, but also how it is distributed among demographic groups. Similarly, in politics observers may want to know not only which candidate has stronger support from the electorate, but also how that support varies by geography, income, and race/ethnicity.

This type of analysis poses significant challenges to internet-based systems because many of the variables of interest (e.g., demographics) are not publicly observable. Thus, one must build a separate classification model for each variable, for example classifying the demographics of a social media user based on linguistic and social evidence. Traditional supervised approaches to this problem suffer from two primary limitations: (1) it is costly and time-consuming to annotate data for each variable of interest for training and validation; and (2) in streaming settings models quickly becoming outdated due to rapidly changing linguistic patterns. For example, Figure 1 shows the association of the term ‘#hillary2016’ on Twitter with various class labels (described in more detail below). We can see that this term was highly indicative of some demographic classes (e.g. college graduates) for almost five months and faded after that. Thus, models need to be robust to rapidly shifting distributions in the data.

To address these challenges, in this paper we propose an approach based on Learning from Label Proportions (LLP) [4]–[7]. Unlike traditional supervised learning, LLP models do not require instance-level annotations for training. Instead, in LLP the training data consist of bags of instances annotated with label proportions – e.g., a collection of 1,000 users, of which 80% are expected to be male. LLP models are appealing in this domain because there are many pre-existing data sources that can provide approximate label proportions. For example, by combining county-level demographics with geolocated tweets, we can associate bags of users with expected demographic distributions. To deal with data drift, we retrain the LLP models daily, which is possible because no additional labeled data is required.

In this paper, we develop an LLP approach to estimate the relationship between political sentiment and demographics during the 2016 U.S. presidential election. We collect 88M geo-tagged tweets posted in the year leading up to election day and group them into 424 counties per day. We use U.S. Census county population data as the expected demographic label proportions. We also use national and state polls (Clin-

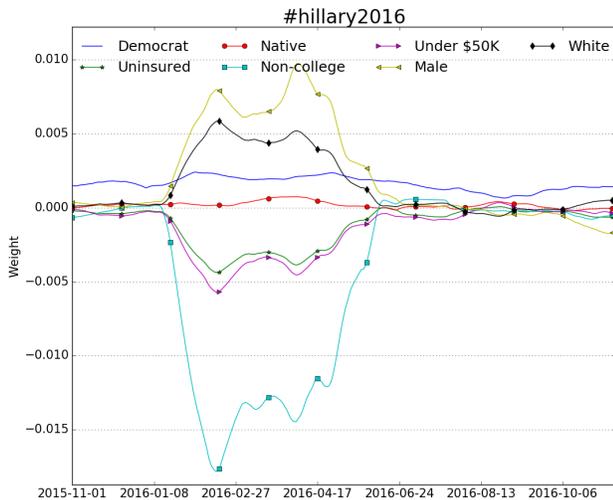


Fig. 1. This figure shows the comparison of classifier coefficients for the term ‘hillary2016’ for different classes. The higher weight shows a strong indication for the given class, and the lower weight shows a substantial evidence of the opposite class

ton vs. Trump) and Cook partisan voting index (PVI)<sup>1</sup> as expected state level label proportions for political sentiment classification. Using these data, we ultimately fit LLP models to classify tweets along seven dimensions: political sentiment (pro-Clinton or pro-Trump), race/ethnicity, education, income, gender, native or foreign born, and health insurance coverage status.

To do so, we propose a new LLP training algorithm, called *weighted label regularization*, which is appropriate for settings in which bags are organized hierarchically — e.g., county bags are nested inside of state bags, which are in turn nested inside of a nation-wide bag. The approach combines ideas from label regularization ([8], [9]) and ridge regression into a scalable model that can be retrained frequently to maintain model freshness. For each day, we retrain all seven LLP models, then calculate conditional probabilities, such as  $P(\text{pro-Clinton} \mid \text{College Graduate})$ . For quantitative evaluation, we compare our estimates to CNN/ORC<sup>2</sup> polls that stratify results by demographics. Additionally, we compare our results on election day with exit polls<sup>3</sup> and the final election results. Our proposed approach produces estimates that closely align with the polls, reducing error by 28-44% over competing baselines on average across all demographic variables.

An additional advantage of our approach is that we can stratify our estimates using combinations of variables, which is often impractical with polling data due to small sample sizes. For example, our estimates of  $P(\text{pro-Clinton} \mid \text{No College and Income} < \$50K)$  show a strong decline in the months prior to the election, in line with journalistic reports of the weakness of Clinton’s support among this group.

<sup>1</sup><http://cookpolitical.com/>

<sup>2</sup><http://orcinternational.com/news-category/cnn-poll/>

<sup>3</sup>A poll of voters taken after they have exited the polling stations.

We also provide some qualitative analysis of how linguistic patterns change over time with respect to political sentiment.

The paper is organized as follows. In Section II, we review related work on internet-based tracking methods and demographic inference, and in Section III we describe the data collected for the experiments. Section IV provides background on LLP models and introduces our proposed approach. In Section V we present our experimental results; Section VI concludes and provides discussion of limitations and future work.

## II. RELATED WORK

Analyzing temporal dynamic of social media is investigated in many recent works. Abel et al. (2010) study temporal dynamic in Twitter for personalized recommendation [10]. For example, their model can detect new users who become interested in a new topic. Yang et al. (2011) develop a spectral clustering algorithm to address how some hashtag’s popularity grows and fades over time [11].

Traditional supervised learning is widely being used in previous works [12]–[14]; however, annotations can be costly to obtain in a timely fashion. Also, many attributes such as political affiliation are hard to interpret, and in a temporal environment such as Twitter, the annotated users became outdated soon. Other work has jointly modeled demographic variables in social networks [15], [16], though again this relies on user-level annotations.

One possible approach to resolving old annotated data is domain adaption. Li et al. (2015) propose the Naive Bayes approach with Expectation Maximization (EM) to predict a new disaster based on labeled data available from Twitter for past catastrophes [17]. The advantage of their model is that they do not need to annotate labeled data for the current disaster with unsupervised domain adaptation. Imran et al. (2016) propose a domain adaptation model for disaster classification [18]. They show that the labels from the previous crisis are useful when the source and the target events are the same types (e.g. earthquakes). They also indicate that cross-language domain adaptation works better when two languages are similar (e.g. Italian and Spanish). However, these methods still base on labeled data on source domain.

Domain adaptation also can be applied to Learning from Label Proportions [19]–[21]. In this approach, the model trained on the domain of origin is used to transfer to the new domain. For social media with temporal dynamic, the model that was fitted previously (e.g. last month), can be transferred to another time (e.g. now) with self-training. The main problem of self-training is scalability and sensitivity to hyperparameters, and it can degrade adaptation to the temporal dynamic of social media.

An attractive alternative is training LLP classifiers with a sliding temporal window. In this case, we do not need domain adaptation, and we can smooth the output of the classifier with moving average to make it more robust to noise. While LLP has a satisfactory result for many classification applications in social science [5], fraud detection [22], and computer vision

[23], [24], to the best of our knowledge, it has not been applied to time series tasks.

The main challenges to using LLP on time series are scalability and robustness to noisy environment of social media (e.g. Twitter). Prior work has proposed an exhaustive greedy bag selection algorithm to deal with noise [25]. While this method has accurate result on some domains, it is not scalable and cannot apply to time series environments. Therefore, a scalable model is required to use in this area.

In this work, we develop a scalable LLP model by using a sliding window for training and estimate the conditional probability between different latent attributes. To improve robustness against inherent noise in social media, we apply moving average instead of using exhaustive bag selection algorithms.

In this paper, we propose the Weighted Label Regularization (WLR) model with several key differences from Label Regularization (LR) for LLP settings. With these contributions, WLR can efficiently be applied to time series data over Twitter with millions of samples. The differences between the two models are as following:

- 1) LR uses softmax, but WLR uses logistic function (because we need only binary classification).
- 2) LR assumes all unlabeled samples have the same weight, while WLR supports weighted samples.
- 3) In WLR model, feature vectors are the average of features of sub bags, and as a result, training is significantly faster than LR, making it feasible to apply to big data.

### III. DATA

For purposes of this study, we collect both individual level data (from social media) and population level data. These data are used to train the LLP models and to make inferences for different social media activity. This section describes the detail of our data collection.<sup>4</sup>

#### A. Twitter data

To understand temporal dynamics in social media, we use the Twitter Streaming API to collect a random sample of geolocated tweets in the United States for roughly one year (Oct 20, 2015, to Nov 7, 2016). We use reverse geocoding to find the originating U.S. county based on the geo-tagged attribute of tweets, and remove tweets which reverse geocoding is failed. After this process, 88M tweets remains with 1.3B tokens and 9M terms. To reduce model complexity, we use only the top 17.5K of the most common unique unigrams that roughly appear in at least 5K tweets.

Then, we create daily bags for each county. Since less populous counties usually lean toward the Republican Party and removing them can introduce bias towards the Democratic Party, to reduce variance, we collapse counties with fewer than 40 tweets per day in each state together. As a result, we totally create 424 county bags (including collapsed bags) per day.

<sup>4</sup>Replication code and data will be made available upon publication.

#### B. Population-level data

We do not have any Twitter labeled data, and the aim of LLP is to use population-level data as a light supervision to predict individual level data. As a result, we need to collect aggregated data as described in this section. Even though it is generally accepted that social media users are not a representative sample of population, an advantage of LLP algorithms is that they are often robust to slight mispecifications of bag proportions [21].

1) *US Census*: For demographic attributes, we collect the latest (2014) county statistics from the U.S. Census. For purposes of this study, we only considered binary classification for 6 attributes: **race** (white or non-white), **education** (college graduate or non-college graduate), **income** (under \$50K household income or \$50K or more), **gender** (male or female), **nativity** (native or foreign born), and **health insurance** (insured or uninsured).

2) *Polls*: To estimate temporal dynamics of political sentiment, we use averaged poll data from the “Real Clear Politics”<sup>5</sup> website. This site reports the moving average of polls from highly graded pollsters for both national and state level, and we use their daily average estimates for Clinton vs. Trump as population-level data for political sentiment classification.

3) *Cook partisan voting index (PVI)*: Cook Political Report periodically reports this index as an estimate of how strongly a state leans toward major parties. For example, the PVI of Florida is “R+2” for 2014, that means Florida tipped 2% more than national average toward Republican Party. We use the latest index (2014<sup>6</sup>) as an additional aggregation level estimate (described in more detail below).

#### C. Data quality

The advantage of the above data is that we can easily associate bags of tweets with label proportions obtained from pre-existing census and polling data. Of course, this data, while convenient, is far from perfect. First, there is selection bias from the fact that Twitter users are not representative of the overall population. Second, census statistics and polling data are themselves only approximations, and so any errors in them will propagate to the trained model. Third, relying on geolocated data presents further challenges to sample size and quality. Despite these challenges, there is considerable evidence in prior work that LLP models are quite robust to noisy label proportions and biased data [8], [9], [21]. This is in part due to the “softness” of the training objective, which accommodates mislabeled instances, and in part due to the fact that the model contains intercept terms that can account for some of the selection bias. (E.g., if younger users are overrepresented in Twitter, the intercept for the age classifier adjusts for this.)

<sup>5</sup><http://www.realclearpolitics.com/>

<sup>6</sup>[https://en.wikipedia.org/wiki/Cook\\_partisan\\_voting\\_index](https://en.wikipedia.org/wiki/Cook_partisan_voting_index)

## IV. MODELS

In this section, we investigate a linear model and propose a non-linear model for learning from label proportion (LLP). Both of these models are scalable and robust for big data settings.

### A. Linear model

We begin with a baseline model that uses ridge regression for LLP. Let  $T_i \in T$  indicate a set of tweets assigned to bag  $i$ , where tweet  $j$  is represented by a  $d$ -dimensional term frequency vector  $x_{ij} \in \mathbb{R}^d$ . The linear model averages the feature vectors for each user in bag  $i$ , and minimizes the mean squared error between the true and predicted label proportions. Let  $\bar{X}_i = \frac{\sum_j x_{ij}}{|T_i|}$  be the average feature vector for each user in bag  $i$ , and let  $\tilde{y}_i$  be the known label proportion for bag  $i$  (e.g., the proportion of males in county  $i$ ). The linear model is simply the dot product between the average feature vector and the model parameters  $\theta \in \mathbb{R}^d$ :

$$h_i = \bar{X}_i^T \theta$$

The  $\theta$  parameters are optimized to minimize mean-squared error with L2 regularization:

$$\theta^* \leftarrow \operatorname{argmin}_{\theta} \frac{1}{|T|} \sum_i (\tilde{y}_i - h_i)^2 + \frac{\lambda}{2} \|\theta\|^2$$

where  $\lambda$  controls the regularization strength.

While this linear model is conceptually simple, recent research has found that it produces accuracy comparable to traditional supervised models on social media tasks [26]. The main advantage of ridge regression for LLP is that it only needs term frequency per bag, without using individual features of samples. This can significantly speed up the training time with accuracy competitive with supervised models such as logistic regression. As a result, it can apply for big data or streaming data, and only the average of features per bag need to be stored in memory.

We use this model for demographic attribute prediction as follows. For each day, we use the mean of features for the prior week's tweets per bag to create feature matrix  $X$  (each row is the average term frequency of one county), and our target variable ( $\tilde{y}$ ) is the normalized population for the corresponding demographic attribute. Since all our demographic attributes are binary, we just need to train ridge regression for one of the classes. For example, for gender classification we compute the proportion of men in each county as the  $\tilde{y}$  vector, and train ridge regression for  $(X, y)$  to optimize  $\theta$ .

To predict the class label for an individual tweet with feature vector  $x$ , we estimate the probability that sample  $x$  is Male as  $x^T \theta$  (truncated between 0 and 1). As a result, if  $x^T \theta > .5$ , we classify this sample as a male, otherwise as female. Also, we use same L2 regularization strength  $\lambda$  for our all experiments, and we find that the results are not very sensitive to this parameter.

Because of reported high accuracy and scalability of ridge regression [26], we use it as a state of the art baseline

model. However, because our population level data for political sentiment classification is at the state level, and our bags are at the county level, we use the label proportion of a state as the corresponding county-level label proportion. Also, we use the sample rate for each county based on the number of samples in that county. We combine polls and PVI index as described in Section IV-B1 to assign label proportions to bags. We call the resulting model **Ridge-NP**.

### B. Non-linear model

Label regularization [8] is a semi-supervised non-linear model (with logistic hypothesis) and is similar to logistic regression for the supervised part. For the semi-supervised part, the model tries to minimize the cross-entropy between the given label proportion and posterior probability estimate of unlabeled data. The original experiments using label regularization assumed that there is a set of labeled data and only one bag of unlabeled data with known label proportion. However, subsequent work has extended the model to multiple unlabeled bags and without any labeled data (i.e., LLP settings) [25].

Scaling label regularization is challenging because at training time it must iterate over each individual instance (tweet) to compute the gradient. Therefore, its training time is much slower than the ridge regression model in the previous section, which only considers a single average feature vectors per bag. We omit label regularization from our experiments below because it is not scalable and requires prohibitively long training time for the millions of training instances in our data; it is also quite sensitive to hyper-parameters. This is the motivation for the present work. We propose a lightweight generalization of label regularization that can use term frequency of county bags. Thus, the training time is much faster than label regularization, allowing us to scale to the current problem domain. We named this scalable model *weighted label regularization (WLR)*.

Let  $X_{u,i}$ ,  $w_{u,i}$ , and  $h_{u,i}$  be the term frequency vector, number of tweets, and the hypothesis for county  $u$  in state  $i$ , and  $\tilde{y}_i$  be the known label proportion for state  $i$ . We define  $h_{u,i}$  same as logistic regression, i.e.

$$h_{u,i} = \sigma(X_{u,i}^T \theta) \quad (1)$$

where  $\theta$  is the model parameter and  $\sigma$  is the logistic function. We define  $\bar{h}_i$  as weighted average of  $h_{u,i}$ :

$$\bar{h}_i = \frac{\sum_u w_{u,i} h_{u,i}}{\sum_u w_{u,i}} \quad (2)$$

Thus, in weighted label regularization, we estimate the state-level proportions as a weighted average of the predicted county-level proportions. Similar to label regularization, we use cross-entropy ( $H$ ) as the error function:

$$\begin{aligned} J(\theta) &= \sum_i H(\tilde{y}_i, \bar{h}_i) + \frac{\lambda}{2} \|\theta\|^2 \\ &= - \sum_i (\tilde{y}_i \log \bar{h}_i + (1 - \tilde{y}_i) \log(1 - \bar{h}_i)) + \frac{\lambda}{2} \|\theta\|^2 \end{aligned} \quad (3)$$

TABLE I  
POPULATION DATA THAT BEING USED FOR EACH MODEL.

Model name	State polls	National polls	PVI
<b>WLR-NP</b>	no	yes	yes
<b>WLR-SN</b>	yes	yes	no
<b>WLR-SNP</b>	yes	yes	yes

where  $\lambda$  is the L2 regularization strength. Our experimental results (Table IV) show that the model is not very sensitive to  $\lambda$ . We set  $\lambda = .01$  for all our other experiments.

We also need the gradient of the cost function to apply the gradient descent algorithm. To do that, we use the gradient of logistic function, i.e.

$$\frac{\partial}{\partial \theta} \sigma(f) = \sigma(f)(1 - \sigma(f)) \frac{\partial}{\partial \theta} f \quad (4)$$

Now, the gradient of the cross-entropy part of the cost function is:

$$\begin{aligned} &= - \sum_i (\tilde{y}_i \frac{\partial}{\partial \theta} \log \bar{h}_i + (1 - \tilde{y}_i) \frac{\partial}{\partial \theta} \log(1 - \bar{h}_i)) \\ &= - \sum_i \left( \frac{\tilde{y}_i}{\bar{h}_i} \frac{\partial \bar{h}_i}{\partial \theta} - \frac{1 - \tilde{y}_i}{1 - \bar{h}_i} \frac{\partial \bar{h}_i}{\partial \theta} \right) \\ &= \sum_i \frac{\bar{h}_i - \tilde{y}_i}{\bar{h}_i(1 - \bar{h}_i)} \frac{\partial \bar{h}_i}{\partial \theta} \\ &= \sum_i \frac{\bar{h}_i - \tilde{y}_i}{\bar{h}_i(1 - \bar{h}_i)} \frac{\partial}{\partial \theta} \frac{\sum_u w_{u,i} h_{u,i}}{\sum_u w_{u,i}} \\ &= \sum_i \frac{\bar{h}_i - \tilde{y}_i}{\bar{h}_i(1 - \bar{h}_i) \sum_u w_{u,i}} \sum_u w_{u,i} h_{u,i} (1 - h_{u,i}) X_{u,i} \\ &= \sum_{u,i} \frac{w_{u,i} h_{u,i} (1 - h_{u,i}) (\bar{h}_i - \tilde{y}_i)}{\bar{h}_i(1 - \bar{h}_i) \sum_u w_{u,i}} X_{u,i} \end{aligned} \quad (5)$$

Finally, any gradient descent algorithm can be used to find parameters. (Although the objective is non-convex, convex optimization has been shown to work well in prior LLP studies [27].) In this study, we use the L-BFGS algorithm [28] to find coefficient  $\theta$  that minimizes the cost function.

To apply WLR for political sentiment training, same as demographic training, we use the average of features for last week per county, and group counties together to create state bags. We also need state label proportions ( $\tilde{y}_i$ ). Because some states are polled more frequently than others, we consider three strategies to assign the label proportion for each state bag for training, summarized in Table I and described below.

1) *WLR-NP*: In this approach, we use the average national poll plus PVI index for all states to assign a label proportion to each state. For example, on Nov 1, 2016, according to the Real Clear Politics site, Clinton polled at 47.5%, and Trump polled at 45.3%. Since we use binary classification, we do not consider third-party candidates. As a result, we estimate the proportion of positive Democratic sentiment as  $47.5/(47.5 + 45.3) = 51.2\%$  at the national level. We use PVI to generate label proportions for each state. For example, Florida has PVI ‘R+2’, so we assign the label proportion for

TABLE II  
THE MAE BETWEEN OUR MODELS AND CNN/ORC POLLS.

Demographic	Ridge-NP	WLR-NP	WLR-SN	WLR-SNP	MOE
US	2.4	1.9	1.9	2	3.3
Midwest	5.7	4.9	5.8	5.2	7
Northeast	3.8	2.5	3.9	3	7
South	3.8	3.3	4.6	3.6	5.6
West	3.6	4.1	3.4	4.3	7
Man	6.7	6.3	7.8	6.2	4.6
Woman	4.9	3.5	5	3.7	4.5
White	12.7	8.5	13.3	8.4	3.7
Non-White	22.3	17.4	21.7	17.8	7.2
Under \$50K	3.4	2.7	3.2	2.2	5.7
\$50K or more	6.3	3.4	5.6	3.8	4.5
College Grad	4.2	1.8	4.8	2.1	4.9
Non-college	4.0	3.1	3.8	3.3	4.5
White college	4.3	3.2	6.1	3.4	5.4
White non-college	16.2	6.4	10.1	6.4	5.1
<b>Average</b>	<b>7.0</b>	<b>4.9</b>	<b>6.7</b>	<b>5</b>	<b>5.3</b>

positive Democratic sentiment in Florida on Nov 1, 2016 as  $51.2\% - 2\% = 49.2\%$ .

2) *WLR-SN*: In this method, rather than using PVI, we restrict the training data to the normalized state polls for states with a poll available on the corresponding day, removing states without polls from the training data for that day.

3) *WLR-SNR*: Similar to the prior method, we use normalized state polls when available. We additionally augment this data using the PVI method above for other states that do not have a poll available on a given day.

### C. Training steps

To apply the proposed models, several preprocessing steps are required. For higher performance, these actions can be sped up by pre-computing steps (such as reverse geocoding and storing the daily average of features for county bags). The primary steps of training at day  $d$  are as follows:

- Add tweets from last week to the training set. Formally, we select all tweets in  $[d-7, d]$  to create the training set.
- Use reverse geocoding to create county bags for training data.
- Tokenize tweets; we remove mentions and URLs and maintain hashtags and description field.
- Compute the average feature vector for each county bag.
- Finally, train LLP models to find model parameters. We use the same hyper-parameters (i.e., random initialization, number of BFGS iterations, and L2 regularization strength) for all experiments.

The reason we retrain every day is to ensure that the model coefficients best reflect the most recent data distribution. While census demographics do not change much, those who participate in political discussions on Twitter on a given day can change rapidly over time, as do the topics that they discuss. Retraining allows the model to capture these latest trends.

### D. Estimating conditional probabilities

In this section, we describe how to infer joint probability (and therefore conditional probability) distributions for different classes for day  $d$ , using our trained models. Let  $B$  be a boundary (e.g. state, county, city) that we are interested in (in

TABLE III  
ESTIMATED PROBABILITY OF VOTING DEMOCRATIC COMPARED TO EXIT POLL.

Demographic	exit-poll	Ridge-NP	WLR-NP	WLR-SN	WLR-SNP
Man	43.6	50.5	48.3	49.8	49
Woman	56.2	53.4	52	54.9	53.1
White	38.9	51.4	42.3	47.5	43.1
Non-White	77.9	53.6	63.1	61	64.4
Under \$50K	55.9	50.4	50.4	53.4	51.5
\$50K or more	49	53.6	50.6	52.4	51.5
College	54.7	54.4	58	59.3	59.1
Non-college	45.8	49	39.5	43.3	40.2
White college	47.9	53.9	51.2	54.8	52.1
White non-col.	28.9	48.4	32	39.1	32.6
Native	47.4	52.2	49.4	52.6	50.4
Foreign born	67.4	52.6	60.6	54.9	61.7
<b>Error</b>		8.8	4.9	6.8	5

this study we use only state boundaries), and  $T_{B,d}$  is the set of all sampled tweets originating from this boundary at day  $d$ . This set can be quickly populated by reverse geocoding. For the sake of simplicity, suppose we are interested in estimating  $P(\text{Democratic, male}|B, d)$ . We propose two methods for this estimation.

**Hard-voting:** In this approach, we compute the number of tweets classified as both ‘Democratic’ and ‘Male’ in  $T_{B,d}$ , and divide that by the number of total tweets in  $T_{B,d}$ . More formally, let  $\theta_D$  and  $\theta_M$  be model parameters for ‘Democratic’ and ‘Male’ class. We estimate the joint probability  $P(D, M|B, d)$  as follows:

$$\frac{|\{x \in T_{B,d} | P(D|x, \theta_D) > .5 \wedge P(M|x, \theta_M) > .5\}|}{|T_{B,d}|} \quad (6)$$

**Soft-voting:** In this method, instead of computing the majority vote for both classes, we compute the average of  $P(D, M|x, \theta_D, \theta_M)$  assuming ‘Democratic’ and ‘Male’ classes are independent (since they are computed independently). More formally, we estimate  $P(D, M|B, d)$  as:

$$\frac{1}{|T_{B,d}|} \sum_{x \in T_{B,d}} P(D|x, \theta_D) P(M|x, \theta_M) \quad (7)$$

In practice, according to Figure 3, we find that for regional attributes (e.g. ‘Midwest’, ‘Florida’) soft-voting works better, and for demographic attributes (e.g. ‘White’) using the weighted average of soft-voting and hard-voting (75% soft, 25% hard) has the best result, and we use this method for our experiments in the next section. Once we have computed the joint probability, we then use the chain rule of probability to compute the desired conditionals; e.g.,  $P(D | M, B, d) = \frac{P(D, M|B, d)}{P(M|B, d)}$ .

## V. RESULTS

We compare the estimates produced by our models both with tracking polls in the year prior to the election and also with exit polls the day of the election. We investigate three research questions:

- **RQ1** Can a model trained on population level data produce accurate estimates of the political sentiment of demographic groups over time?

- **RQ2** What is the relative impact of the different methods of assigning label proportions to bags (i.e., methods WLR-NP, WLR-SN, and WLR-SNP above)?
- **RQ3** How do certain terms change over time with respect to their association with demographics and political sentiment?

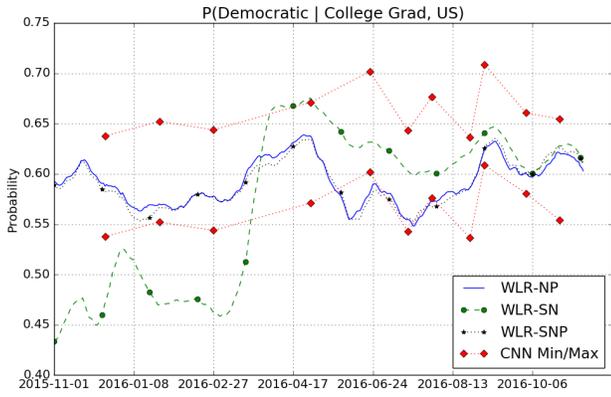
In the first experiment, we estimate conditional probabilities of political sentiment given demographic classes at the national level by computing the weighted average of the state-level estimates. We compare these estimates with the demographic breakdown of polls from CNN/ORC. There were 11 CNN/ORC polls conducted during this time, with five regions (US, Midwest, Northeast, South, and West), and ten demographic breakdown attributes. Table II shows the mean absolute error (MAE) between model prediction and CNN/ORC result. The last column in this table shows the average margin of error (MOE) of polls. For all but four demographic classes, **WLR-NP** has an error rate less than the margin of error. The largest error belongs to race classes, which is in line with previous work [26]. According to this table, **WLR-NP** is more accurate than **WLR-SNP**. Also, **WLR-SN** and **Ridge-NP** have an error rate above the margin of error.

The lowest error in Table II belongs to ‘College grad’ and ‘Under \$50K’ for demographic breakdowns. Figure 2a plots the daily probability of being Democratic (smoothed with 14 days moving average) given college graduate, compared to CNN/ORC lowest and highest margin of error. According to this plot, **WLR-NP** is close to the **WLR-SNP** method, and except for one poll, it is within the margin of CNN/ORC poll error. Furthermore, it shows that **WLR-SN** has the highest error. Figure 2b plots the same pattern for ‘Under \$50K’ class and shows that **WLR-NP** has the lowest error; except for one poll, it is in the margin of CNN/ORC poll error. Finally, Figure 2c shows the probability of voting Democratic given both education and income level. Here, we do not have access to polls reporting this combination of variables, in part because small sample sizes make these difficult to estimate using traditional polling methods. However, we note a significant drop in Democratic support among people with low income and low education levels in late August/early September prior to the election.

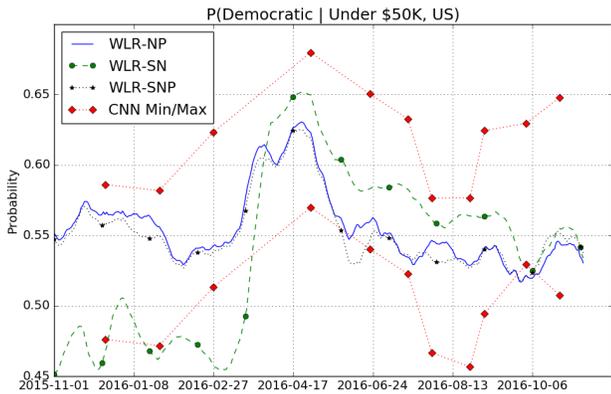
In the next experiment, we use the exit poll results and compare them with our model predictions generated one day before the election date. Table III compares our models with exit polls. Again, **WLR-NP** has the lowest error rate, and ‘Non-white’ class has the highest error. According to this table, ‘\$50K or more’, ‘White non-college graduate’, and ‘Native’ classes have the lowest error.

To show the sensitivity of model parameters, we run our best model (**WLR-NP**) with different L2 regularization strength ( $\lambda$ ). Table IV shows the error rate of **WLR-NP** with various model parameters. According to this table, the error rate is stable for small values of  $\lambda$ . This result shows that the model is not very sensitive to L2 regularization.

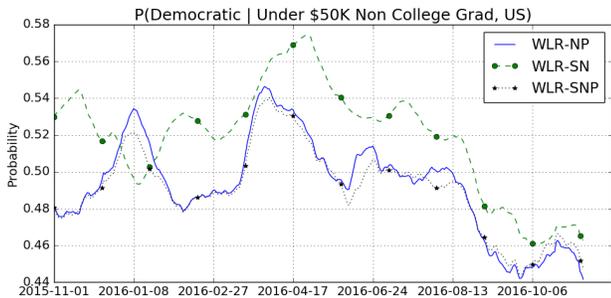
While our primary goal is not to predict election results,



(a) college



(b) income



(c) college and income

Fig. 2. Model predictions for the probability of pro-Clinton sentiment, conditioned on (a) education level, (b) income level, (c) combination of education and income level. While (a) and (b) plot the CNN polls for comparison, no poll is available for (c), since traditional polls typically do not report multiple demographic splits due to small sample sizes.

as an additional validation measure we also compare our predictions (at one day before election day) to election results. Table V compares our prediction of being Democratic with the election result for battleground states. While all models have a very similar average error rate, **WLR-NP** incorrectly predicts the winner of only 5 states (Colorado, Iowa, Michigan, Pennsylvania, Virginia), the fewest among all approaches. These results suggest that models based on state polls (**WLR-SN** and **WLR-SNP**) have a poor prediction. This may in part

TABLE IV  
EFFECT OF L2 REGULARIZATION STRENGTH ON ERROR.

Lambda	CNN/ORC	Exit poll
.0001	4.9	4.8
.001	4.9	4.8
.01	4.9	4.9
.1	5.1	5
1	5.6	5.6

TABLE V  
LIKELIHOOD OF BEING DEMOCRATIC FOR BATTLEGROUND STATES COMPARE TO ELECTION RESULTS.

State	Truth	WLR-NP	WLR-SN	WLR-SNP
AZ	47.7	43.5	49.4	44.4
CO	51.1	49.3	49.7	49.3
FL	49.3	49.8	50	51.4
GA	47.1	48.1	47.6	50.8
IA	44.9	50.1	49	50.6
MI	49.8	51.5	52.2	50.9
MN	50.8	51.4	53.4	52.7
NC	48	47.1	48.4	49.1
NH	50.1	52.6	48.9	51.9
NV	51.3	52.3	48.1	51.4
OH	45.5	45.9	48.4	46.3
PA	49.4	51.6	50.9	52.6
VA	52.6	48.2	51.6	49.8
WI	49.5	49.9	52.8	50.5
Error	0	1.9	1.9	2.2

because of inaccurate polls in some states (notably Florida and the Midwest).

Figure 3 plots the effect of weighted average between soft-voting and hard-voting. This plot shows the error (MAE) of **WLR-NP** using different weighted averages between soft-voting and hard-voting. The leftmost of this plot shows using only hard-voting (0% soft-voting), and the rightmost illustrates the error of using only (100%) soft-voting. In this plot we divide CNN/ORC polls to regional (e.g. ‘Midwest’, ‘Florida’) and demographic (e.g. ‘White’) attributes, and according to this plot for regional attributes (i.e. CNN/ORC regional and election result) soft-voting works better. However, for demographic attributes (i.e. CNN/ORC demographics and exit polls) using the weighted average of soft-voting and hard-voting (75% soft, 25% hard) has the best result.

These results answer our first research question by indicating that the joint probability (and therefore conditional probability) of different latent attributes can be estimated by models trained on population-level data, and we evaluate our models with CNN/ORC polls, exit polls, and election result. Our models show that while some demographic attributes (such as race) are hard to predict, some characteristics such as income and college graduation are easier to predict, and as a result, it affects the accuracy of the conditional probabilities.

To answer our second research question, we present evidence that **WLR-NP** by using national polls and PVI index has the best result, and **WLR-SN** has the worst result due to overfitting state polls. We believe that is because of noise in state polls in some regions. Further studies are required to investigate the source of inaccuracy in these states.

To answer our third research question, we select **WLR-**

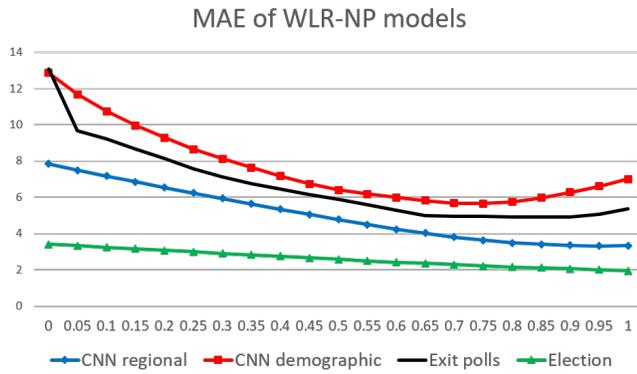


Fig. 3. The MAE of WLR-NP with different weighted averages between soft-voting and hard-voting.

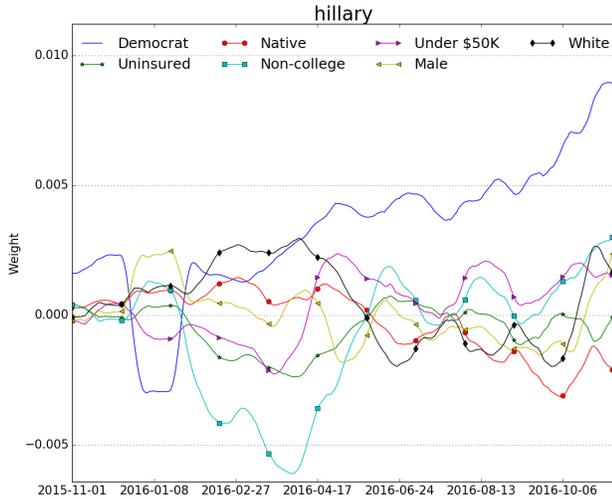


Fig. 4. Weights of term 'hillary' for different classes.

**NP** as the best model for political affiliation prediction and report how using a term can change over time. Figure 4 reports changes of weights (normalized to unit vector) for term 'Hillary' for different classes (smoothed by 30 days moving average). According to this plot, except January, the unigram 'Hillary' has a growing indication for 'Democratic' class. But, for demographic attributes, its sign changes over time. For example, the term is a weak indicator of 'native born' class before July, and after that becomes indicative of the 'foreign born' class. In addition, according to this plot all demographic weights converge to near zero at nomination time, that can be in part because all classes use this term at that point.

On the other hand, according to Figure 5, the term '#trump' has stable indication over one year. Before April, it is almost neutral for all demographic classes and a weak indicator for 'Democratic' class. After April, its indication grows over time to become strongly indicative of 'Republican', 'Non-college graduate' and 'under \$50k' classes. Mildly positive coefficients are found for 'White', 'Uninsured', 'Native', and 'Male' classes.

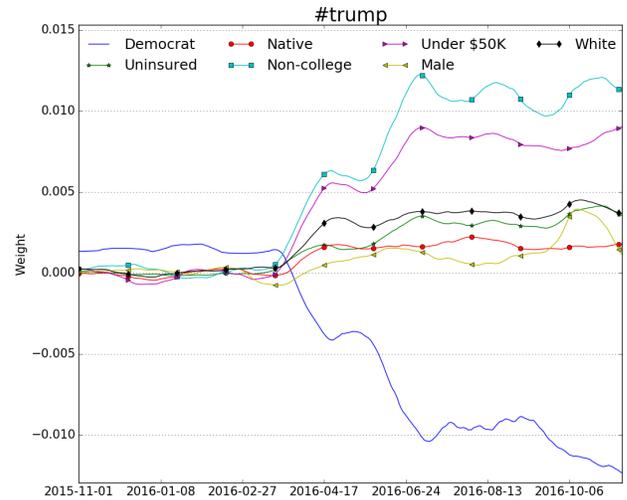


Fig. 5. Weights of term '#trump' for different classes.

Finally, Figure 6 plots weights of some terms for both race and political sentiment classes to show how the unigram indication changes over time (all weights are smoothed with 30 days moving average). For each term, its weight starts from 'x' mark (on Nov 1, 2015) to 'o' mark (on Nov 7, 2016). There are four quartiles in this plot. We select unigrams with the highest indication changes, and some terms (i.e. 'drinking', '#healthcare', 'God', and 'check') keep in one quartile for entire year. For example, the term 'drinking' is an indicator for both 'Democratic' and 'White' classes for the whole year. That is in part because according to 2013 national survey on drug use and health from U.S. Department of Health and Human Services<sup>7</sup>, white Americans use more alcohol than other races/ethnicities, and the rate of alcohol consumption increases with increasing levels of education (which correlates with Democratic political affiliation).

Finally, some terms (i.e. 'international' and 'university') have a solid indication for race classes, but multiple indications for political classes. That is in part because of more temporal dynamics for political classes in election season. Also, the term '#trump' starts from almost neutral and leads to 'Republican' class over time, and becomes a weak indicator for 'white' class.

## VI. CONCLUSIONS AND FUTURE WORK

In conclusion, we found that the population-level data can be used to mine conditional probability of demographic and opinion attributes from Twitter. Our first contribution is scalability compared to previous works. Our proposed model, weighted label regularization, is a scalable generalization of label regularization that can apply to domains where bags of users are grouped into smaller sub bags. The training time of this model is significantly faster than label regularization because it does not require individual features of users in sub

<sup>7</sup><http://www.samhsa.gov/data/sites/default/files/NSDUHresultsPDFWHTML2013/Web/NSDUHresults2013.pdf>

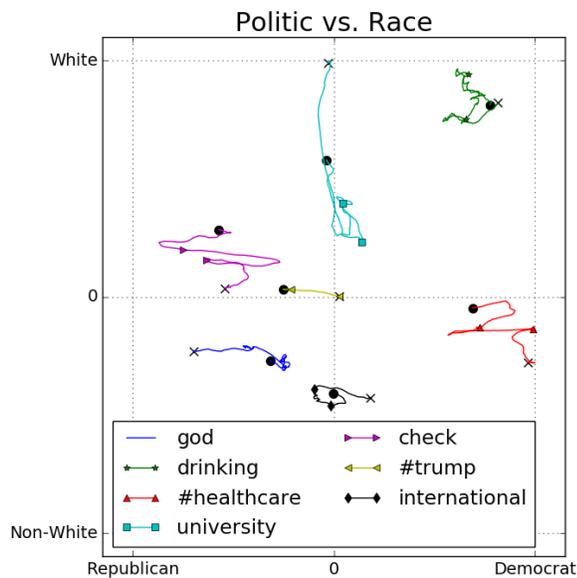


Fig. 6. Term weights for political and race classes.

bags, relying instead on the average of feature values and the number of samples in each sub bag. This difference makes weighted label regularization applicable to the data size in this domain.

Our second contribution is to investigate one step beyond the classification task by estimating joint and conditional probabilities between different latent attributes. This process benefits social scientist to understand the opinion of different demographic populations.

Finally, our experimental results show that using national polls with PVI has the lowest error and some state polls appear to be inaccurate. This method, in turn, can be utilized as a supplement to polls to discover public opinion for election candidates.

In the future, we will investigate new models to track opinion and public health in domains with high temporal dynamics and propose more scalable and accurate models to adapt to these dynamics.

#### ACKNOWLEDGMENT

This research was funded in part by the National Science Foundation under grants #IIS-1526674 and #IIS-1618244.

#### REFERENCES

- [1] M. Dredze, "How social media will change public health," *IEEE Intelligent Systems*, vol. 27, no. 4, pp. 81–84, 2012.
- [2] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From Tweets to polls: Linking text sentiment to public opinion time series," in *International AAAI Conference on Weblogs and Social Media*, Washington, D.C., 2010.
- [3] S. Gopinath, J. S. Thomas, and L. Krishnamurthi, "Investigating the relationship between the content of online word of mouth, advertising, and brand performance," *Marketing Science*, vol. 33, no. 2, pp. 241–258, 2014.

- [4] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le, "Estimating labels from label proportions," *Journal of Machine Learning Research*, vol. 10, pp. 2349–2374, 2009. [Online]. Available: <http://jmlr.org/papers/volume10/quadrianto09a/quadrianto09a.pdf>
- [5] F. X. Yu, S. Kumar, T. Jebara, and S. Chang, "On learning with label proportions," *CoRR*, vol. abs/1402.5902, 2014. [Online]. Available: <http://arxiv.org/abs/1402.5902>
- [6] R. E. Schapire, M. Rochery, M. G. Rahim, and N. K. Gupta, "Incorporating prior knowledge into boosting," in *Proceedings of the Nineteenth International Conference*, 2002, pp. 538–545.
- [7] R. Jin and Y. Liu, "A framework for incorporating class priors into discriminative classification," in *In PAKDD*, 2005.
- [8] G. S. Mann and A. McCallum, "Simple, robust, scalable semi-supervised learning via expectation regularization," in *Proceedings of the 24th International Conference on Machine Learning*, ser. ICML '07. New York, NY, USA: ACM, 2007, p. 593600. [Online]. Available: <http://doi.acm.org/10.1145/1273496.1273571>
- [9] J. Graça, K. Ganchev, and B. Taskar, "Expectation maximization and posterior constraints," in *NIPS*, vol. 20, 2007, pp. 569–576.
- [10] F. Abel, Q. Gao, G.-J. Houben, and K. Tao, "Analyzing Temporal Dynamics in Twitter Profiles for Personalized Recommendations in the Social Web," in *Proceedings of ACM WebSci '11, 3rd International Conference on Web Science, Koblenz, Germany*. ACM, June 2011.
- [11] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 177–186. [Online]. Available: <http://doi.acm.org/10.1145/1935826.1935863>
- [12] M. Pennacchiotti and A.-M. Popescu, "A machine learning approach to twitter user classification," in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011. [Online]. Available: <http://dblp.uni-trier.de/db/conf/icwsm/icwsm2011.html>
- [13] R. Cohen and D. Ruths, "Classifying political orientation on twitter: It's not easy!" in *ICWSM*, 2013.
- [14] E. Colleoni, A. Rozza, and A. Arvidsson, "Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data," *Journal of Communication*, vol. 64, no. 2, pp. 317–332, 2014.
- [15] D. Chakrabarti, S. Funiak, J. Chang, and S. Macskassy, "Joint inference of multiple label types in large networks," in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 874–882.
- [16] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 15–24.
- [17] H. Li, N. Guevara, N. Herndon, D. Caragea, K. Neppalli, C. Caragea, A. C. Squicciarini, and A. H. Tapia, "Twitter mining for disaster response: A domain adaptation approach," in *12th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Krystiansand, Norway, May 24-27, 2015*, L. Palen, M. Büscher, T. Comes, and A. L. Hughes, Eds. ISCRAM Association, 2015. [Online]. Available: [http://idl.iscram.org/files/hongminli/2015/1234\\_HongminLi\\_et al2015.pdf](http://idl.iscram.org/files/hongminli/2015/1234_HongminLi_et al2015.pdf)
- [18] M. Imran, P. Mitra, and J. Srivastava, "Cross-language domain adaptation for classifying crisis-related short messages," *CoRR*, vol. abs/1602.05388, 2016. [Online]. Available: <http://arxiv.org/abs/1602.05388>
- [19] J. Margolis and A. Fisher, *Unlocking the clubhouse: Women in computing*. MIT press, 2003.
- [20] C. Kadar and J. Iria, "Domain adaptation for text categorization by feature labeling," in *ECIR*, 2011, pp. 424–435. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1996889.1996944>
- [21] E. M. Ardehaly and A. Culotta, "Domain adaptation for learning from label proportions using self-training," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, 2016, pp. 3670–3676. [Online]. Available: <http://www.ijcai.org/Abstract/16/516>
- [22] S. Rping, "Svm classifier estimation from group probabilities," 2010.
- [23] H. Kück and N. de Freitas, "Learning about individuals from group statistics," *CoRR*, vol. abs/1207.1393, 2012. [Online]. Available: <http://arxiv.org/abs/1207.1393>
- [24] K.-T. Lai, F. X. Yu, M.-S. Chen, and S.-F. Chang, "Video event detection by inferring temporal instance labels," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR

- '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 2251–2258. [Online]. Available: <http://dx.doi.org/10.1109/CVPR.2014.288>
- [25] E. Mohammady Ardehaly and A. Culotta, “Inferring latent attributes of twitter users with label regularization,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–June 2015, pp. 185–195. [Online]. Available: <http://www.aclweb.org/anthology/N15-1019>
- [26] E. Mohammady and A. Culotta, “Using county demographics to infer attributes of twitter users,” in *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, 2014.
- [27] G. S. Mann and A. McCallum, “Generalized expectation criteria for semi-supervised learning with weakly labeled data,” *J. Mach. Learn. Res.*, vol. 11, p. 955984, Mar. 2010. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1756006.1756038>
- [28] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, “A limited memory algorithm for bound constrained optimization,” *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.