Inferring latent attributes of Twitter users with label regularization

Ehsan Mohammady Ardehaly and **Aron Culotta**
Department of Computer Science
Illinois Institute of Technology
Chicago, IL 60616
emohamm1@hawk.iit.edu, aculotta@iit.edu

Abstract

Inferring latent attributes of online users has many applications in public health, politics, and marketing. Most existing approaches rely on supervised learning algorithms, which require manual data annotation and therefore are costly to develop and adapt over time. In this paper, we propose a lightly supervised approach based on label regularization to infer the age, ethnicity, and political orientation of Twitter users. Our approach learns from a heterogeneous collection of soft constraints derived from Census demographics, trends in baby names, and Twitter accounts that are emblematic of class labels. To counteract the imprecision of such constraints, we compare several constraint selection algorithms that optimize classification accuracy on a tuning set. We find that using no user-annotated data, our approach is within 2% of a fully supervised baseline for three of four tasks. Using a small set of labeled data for tuning further improves accuracy on all tasks.

1 Introduction

Data annotation is a key bottleneck in applying supervised machine learning to language processing problems. This is especially problematic in streaming settings such as social media, where models quickly become dated as new linguistic patterns emerge. An attractive alternative is *lightly supervised learning* (Schapire et al., 2002; Jin and Liu, 2005; Chang et al., 2007; Graça et al., 2007; Quadrianto et al., 2009; Mann and McCallum, 2010; Ganchev et al., 2010). In this approach, classifiers are trained from a set of domain-specific *soft* constraints, rather than individually labeled instances. For example, *label regularization* (Mann and McCallum, 2007; Graça et al., 2007) uses prior knowledge of the expected label distribution to fit a model from large pools of unlabeled instances. Similarly, annotating features with their expected class frequency has proven to be an efficient way of bootstrapping from domain knowledge (Druck et al., 2009; Melville et al., 2009; Settles, 2011).

In this paper we use lightly supervised learning to infer the age, ethnicity, and political orientation of Twitter users. Lightly supervised learning provides a natural method for incorporating the rich, declarative constraints available in social media. Our approach pairs unlabeled Twitter data with constraints from county demographics, trends in first names, and exemplar Twitter accounts strongly associated with a class label.

Prior applications of label regularization use a small number of highly-accurate constraints; for example, Mann and McCallum (2007) use a single constraint that is the true label proportions of an unlabeled dataset, and Ganchev and Das (2013) use cross-lingual constraints from aligned text. In contrast, we use hundreds of constraints that are heterogeneous, overlapping, and noisy. For example, we constrain the predicted attributes of users from a county to match those collected by the Census, despite the known non-representativeness of Twitter users (Mislove et al., 2011). Furthermore, users from that county who list first names in their profile have additional constraints imposed upon them, which may conflict with the county constraints.

Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pages 185–195,
Denver, Colorado, May 31 – June 5, 2015.

To deal with such noisy constraints, we explore forward selection algorithms that choose from hundreds of soft constraints to optimize accuracy on a tuning set. We find that this approach is competitive with a fully supervised approach, with the added advantage of being less reliant on labeled data and therefore easier to update over time. Our primary research questions and answers are as follows:

RQ1. What effect do noisy constraints have on label regularization? We find that simply using all constraints, ignoring noise and overlap, results in surprisingly high accuracy, within 2% of a fully-supervised approach on three of four tasks. For age classification, the constraint noise appears to substantially degrade accuracy.

RQ2. How can we select the most useful constraints? Using a small tuning set, we find that our forward selection algorithms improve label regularization accuracy while using fewer than 10% of the available constraints. Constraint selection improves age classification accuracy by nearly 18% (absolute).

RQ3. Which constraints are most informative? We find that follower constraints result in the highest accuracy in isolation, yet the constraint types appear to be complementary. For three of four tasks, combining all constraint types leads to the highest accuracy.

In the following, we first review related work in lightly supervised learning and latent attribute inference, then describe the Twitter data and constraints. Next, we formalize the label regularization problem and our constraint selection algorithms. Finally, we present empirical results on four classification tasks and conclude with a discussion of future work.

2 Related Work

Inferring demographic attributes of users in social media with supervised learning is a growing area of interest, with applications in public health (Dredze, 2012), politics (O'Connor et al., 2010) and marketing (Gopinath et al., 2014). Attributes considered include age (Nguyen et al., 2011; Al Zamal et al., 2012), ethnicity (Pennacchiotti and Popescu, 2011; Rao et al., 2011), and political orientation (Conover et al., 2011; Barberá, 2013).

The main of drawback supervised learning in social media is that human annotation is expensive and error-prone, and collecting pseudo-labeled data by self-identifying keywords is noisy and biased (e.g., searching for profiles that mention political orientation). For these reasons we investigate lightly-supervised learning, which takes advantage of the plentiful unlabeled data.

Previous work in lightly-supervised learning has developed methods to train classifiers from prior knowledge of label proportions (Jin and Liu, 2005; Chang et al., 2007; Musicant et al., 2007; Mann and McCallum, 2007; Quadrianto et al., 2009; Liang et al., 2009; Ganchev et al., 2010; Mann and McCallum, 2010; Chang et al., 2012; Wang et al., 2012; Zhu et al., 2014) or prior knowledge of features-label associations (Schapire et al., 2002; Haghighi and Klein, 2006; Druck et al., 2008; Melville et al., 2009). In addition to standard document categorization tasks, lightly supervised approaches have been applied to named-entity recognition (Mann and McCallum, 2010; Ganchev and Das, 2013; Wang and Manning, 2014), dependency parsing (Druck et al., 2009; Ganchev et al., 2009), language identification (King and Abney, 2013), and sentiment analysis (Melville et al., 2009).

One similarly-motivated work is that of Chang et al. (2010), who infer race/ethnicity of online users using name and ethnicity distributions provided by the U.S. Census Bureau. This external data is incorporated into the model as a prior; however, no linguistic content is used in the model, limiting the coverage of the resulting approach. Oktay et al. (2014) extend the work of Chang et al. (2010) to also include statistics over first names.

Other work has inferred population-level statistics from social media; e.g., Eisenstein et al. (2011) use geolocated tweets to predict zip-code statistics of demographic attributes of users, and Schwartz et al. (2013) predict county health statistics from Twitter. However, no user-level attributes are predicted.

Patrini et al. (2014) build a Learning with Label Proportions (LLP) model with the objective to learn a supervised classifier when, instead of labels, only label proportions for bags of observations

are known. Their empirical results demonstrate that their algorithms compete with or are just percents of AUC away from the supervised learning approach.

In preliminary work (Mohammady and Culotta, 2014), we fit a regression model to predict the ethnicity distribution of a county based on its Twitter usage, then applied the regression model to classify individual users. In contrast, here we use label regularization, which can more naturally be applied to user-level classification and can incorporate a wider range of constraint types.

3 Data

In this section we describe all data and constraints collected for our experiments.

3.1 Labeled Twitter Data

For validation (and for tuning some of the methods) we annotate Twitter users according to age, ethnicity, and political orientation. We collects four disjoint datasets for this purpose:

Race/ethnicity: This data set comes from the research of Mohammady and Culotta (2014). They categorized 770 Twitter profiles into one of four categories (Asian, Black, Latino, White). They used the Twitter Streaming API to obtain a random sample of 1,000 users, filtered to the United States. These were manually categorized by analyzing the profile, tweets, and profile image for each user, discarding those for which race could not be determined (230/1,000; 23%). The category frequency is Asian (22), Black (263), Latino (158), White (327). For each user, they collected the 200 most recent tweets using the Twitter API. We refer to this dataset as the **race** dataset.

Age: Annotating Twitter users by age can be difficult, since it is rarely explicitly mentioned. Similar to prior work (Rao et al., 2010; Al Zamal et al., 2012), we divide users into those below 25 and those above above 25 years old. Using the idea from Al Zamal et al. (2012), we use the Twitter search API to find tweets with phrases like "happy 30th birthday to me," and then we collect those users and download their 200 most recent tweets using the Twitter API. We collect 1,436 users (771 below 25 and 665 above 25). While this sampling procedure introduces some selection bias, it provides a useful form of validation in the absence of expedient alternatives. We refer to this dataset as the **age** dataset.

Politician: Inspired by works of (Cohen and Ruths, 2013), we select the official Twitter accounts of members of the U.S. Congress. We select 189 Democratic accounts and and 188 Republican accounts and download their most recent 200 tweets. We refer to this dataset as the **politician** dataset.

Politician-follower: As the **politician** dataset is not representative of typical users, we collect a separate political datasets. We first collect a list of followers of the official Twitter accounts for both parties ("thedemocrats" and "gop"). We randomly select 598 likely Democrats and 632 likely Republicans, and download the most recent 200 tweets for each user. While the labels for these data may contain moderate noise (since not everyone who follows "gop" is Republican), a manual inspection did not reveal any mis-annotations. We refer to this as the **politician-follower** dataset.[1]

We split each of the datasets above into 40% tuning/training and 60% testing (though not all methods will use the training set, as we describe below).

3.2 Unlabeled Twitter Data

Label regularization depends on a pool of unlabeled data, along with soft constraints over the label proportions in that data. Since many of our constraints involve location, we use the Twitter streaming API to collect 1% of geolocated tweets, using a bounding box of the United States (48 contiguous states plus Hawaii and Alaska). In order to assign each tweet to a county, we use the U.S. Census' center of population data.[2] We use this data to map each geolocated Twitter user to a corresponding county. We use the k-d tree algorithm (Maneewongvatana and Mount, 2002) to find the nearest center of population for each tweet and use a threshold to discard tweets that are not within a specified distance of any county center. In total, we collect 18 million geolocated tweets from 2.7 million unique users.

[1] We were unfortunately unable to obtain the annotated political data of Cohen and Ruths (2013) for direct comparison.

[2] https://www.census.gov/geo/reference/centersofpop.html

3.3 Constraints

Finally, we describe the soft constraints used by label regularization. Each constraint will apply to a (possibly overlapping) subset of users from the unlabeled Twitter data. For all constraints below, we only include the constraint for consideration if at least 1,000 unlabeled Twitter users are matched. For example, if we only have 500 users from a county, we will not use that county's demographics as a constraint. This is to ensure that there is sufficient unlabeled data for learning. We consider three classes of constraints:

County constraints (cnt): The U.S. Census produces annual estimates of the ethnicity and age demographics for each county. We use the most recent decennial census (2010) to compute the proportion of each county that is below and above 25 years old (to match the labels of the annotated data). We additionally use the 2012 updated estimates of ethnicity by county, restricting to Asian, Black, Latino, and White. Each constraint, then, is applied to the users assigned to that county in the unlabeled data. For example, there are 46K unlabeled users from Cook County, which the Census estimates as 45% White. We consider 3,000 total counties as constraints, of which roughly 500 are retained for consideration after filtering those that match fewer than 1,000 users.

Name constraints (nam): Silver and McCanc (2014) recently demonstrated how a person's first name can often indicate their age. The Social Security Administration reports the frequencies of names given to children born in a given year,[3] and its actuarial tables[4] estimate how many people born in a given year are still alive. From these data, one can estimate the age distribution of people with a given name. For example, the median age of someone named "Brittany" is 23. With this approach, we can assign constraints indicating the fraction of people with a given name that are above and below 25 years old.

For each user in the unlabeled Twitter data, we parse the "name" field of the profile, assuming that the first token represents the first name. Constraints are assigned to users with matching names. We consider more than 50K total name constraints, of which we retain 175 that match a sufficient number of users. For example, there are roughly 1,600 unlabeled users with the first name Katherine; the constraint specifies that 86% of them are under 25.

Follower constraints (fol): Our final type of constraint uses Twitter accounts and hashtags strongly associated with a class label. The constraint applies to users that follow such exemplar accounts or use such hashtags. We consider two sources of such constraints. For age and race, we download demographic data for 1K websites from Quantcast.com, an audience measurement company that tracks the demographics of visitors to millions of websites (Kamerer, 2013). We then identify the Twitter accounts for each website. For example, one constraint indicates that 12% of Twitter users who follow "oprah" are Latino. For political constraints, we manually identify 18 Twitter accounts or hashtags that are strongly associated with either Democrats or Republicans.[5] The constraint specifies that 90% of users that follow one of these accounts (or use one of these hashtags) are affiliated with the corresponding party. (We omit constraints use to construct the labeled data for the **politician-follower** data.)

4 Label Regularization

Our goal is to learn a classification model using the unlabeled Twitter data and the constraints described above. The idea of label regularization is to define an objective function that enforces that the predicted label distribution for a set of unlabeled data closely matches the expected distribution according to a constraint.

We select multinomial logistic regression as our classification model. Given a feature vector x, a class label y, and set of parameter vectors $\theta = \{\theta_{y_1} \dots \theta_{y_k}\}$ (one vector per class), the conditional distribution of y given x is defined as follows:

$$p_\theta(y|x) = \frac{\exp(\theta_y \cdot x)}{\sum_{y'} \exp(\theta_{y'} \cdot x)}$$

[3] http://www.ssa.gov/oact/babynames/

[4] http://www.ssa.gov/oact/NOTES/as120/LifeTables_Tbl_7.html

[5] For Democrats: thedemocrats, wegoted, dccc, collegedems, dennis_kucinich, sensanders, repjohnlewis, keithellison, #p2. For Republicans: gop, nrsc, the_rga, repronpaul, senrandpaul, senmikelee, repjustinamash, gopleader, #tcot

Typically, θ is set to maximize the likelihood of a labeled training set. Instead, we will optimize the objective defined in Mann and McCallum (2007), using only unlabeled data and constraints.

Let $U = \{U_1 \dots U_k\}$ be a set of sets, where U_j consists of unlabeled feature vectors x. The elements of U may be overlapping. Let $\tilde{p}_j$ be the expected label distribution of U_j. E.g., $\tilde{p}_j = \{.9, .1\}$ would indicate that 90% of examples in U_j are expected to have class label 0. The combination of $(U_j, \tilde{p}_j)$ is called a **constraint**.

Our goal, then, is to set θ so that the predicted label distribution matches $\tilde{p}_j$, for all j. Since using the predicted class counts results in an objective that is non-differentiable, Mann and McCallum (2007) instead use the model's posterior distribution:

$$\hat{q}_j(y) = \sum_{x \in U_j} p_\theta(y|x)$$

$$\hat{p}_j(y) = \frac{\hat{q}_j(y)}{\sum_{y'} \hat{q}_j(y')}$$

where $\hat{p}_j$ is the normalized form of $\hat{q}_j$. Then, we want to set θ such that $\hat{p}_j$ and $\tilde{p}_j$ are close. Mann and McCallum (2007) use KL-divergence, which is equivalent to augmenting the likelihood with a Dirichlet prior over expectations where values for the priors are proportional to $\tilde{p}_j$. KL-divergence can be factored into two parts:

$$= -\sum_y \tilde{p}_j(y) \log \hat{p}_j(y) + \sum_y \tilde{p}_j(y) \log \tilde{p}_j(y)$$
$$= H(\tilde{p}_j, \hat{p}_j) - H(\tilde{p}_j)$$

where $H(\tilde{p}_j)$ is constant for each j, and so we need to minimize $H(\tilde{p}_j, \hat{p}_j)$ in order to minimize KL-divergence, where $H(\tilde{p}_j, \hat{p}_j)$ is the cross-entropy of the hypothesized distribution and the expected distribution for U_j.

We additionally use L2 regularization, resulting in our final objective function:

$$J(\theta) = \sum_j H(\tilde{p}_j, \hat{p}_j) + \frac{1}{\lambda} \sum_y ||\theta_y||_2^2$$

In practice we find that λ does not need tuning for each data set. We set it simply to:

$$\lambda = \frac{C}{\sum_j |U_j|}$$

We set C to 1.3e10 in our experiments. Mann and McCallum (2007) compute the gradient of cross-entropy as follows:

$$\frac{\partial}{\partial \theta_k} H(\tilde{p}_j, \hat{p}_j) = -\sum_{x \in U_j} \sum_y p_\theta(y|x) x_k$$
$$\times \left(\frac{\tilde{p}_j(y)}{\hat{p}_j(y)} - \sum_{y'} \frac{\tilde{p}_j(y) \times p_\theta(y'|x)}{\hat{p}_j(y)} \right)$$

The gradient for θ_k is then a sum of the gradients for each constraint j. In order to minimize the objective function, we use gradient descent with L-BFGS (Byrd et al., 1995). (While the objective is not guaranteed to be convex, this approximation has worked well in prior work.) To help reduce overfitting, we use early-stopping (10 iterations).

Temperature: Mann and McCallum (2007) find that sometimes label regularization returns a degenerate solution. For example, for a three class problem with constraint $\tilde{p}_j(y) = \{.5, .35, .15\}$, it may find a solution such that $p_\theta(y) = \{.5, .35, .15\}$ for every instance and as a result all of the instances are assigned the same label. To avoid this behavior Mann and McCallum (2007) introduce a temperature parameter T into the classification function as follows:

$$p_\theta(y|x) = \frac{\exp(\theta_y \cdot x / T)}{\sum_y \exp(\theta_y \cdot x / T)}$$

In practice we find that we can set T to two for binary classification and ten for multi-class problems.

While the approach described above closely follows Mann and McCallum (2007), we note two important distinctions: we use no labeled data in our objective, and we consider a set of hundreds of noisy, overlapping constraints (as opposed to only a handful of precise constraints).

4.1 Constraint Selection

As described above, our proposed constraints are undoubtedly inexact. For example, it is generally accepted that social media users are not a representative sample of the population. E.g., younger, urban and minority populations tend to be overrepresented on Twitter (Mislove et al., 2011; Lenhart and Fox, 2009), and Latino users tend to be underrepresented on Facebook (Watkins, 2009). Thus, it is incorrect to

assume that the demographics of Twitter users from a county match those of all people from a county. While it may be possible to directly adjust for these mismatches using techniques from survey reweighting (Gelman, 2007), it is difficult to precisely quantify the proper weights in this context.

Instead, we propose a search-based approach inspired by feature selection algorithms commonly used in machine learning (Guyon and Elisseeff, 2003). The idea is to select the subset of constraints that result in the most accurate model. We first assume the presence of a small set of labeled data $L = \{(x_1, y_1) \dots (x_n, y_n)\}$. Given a set of constraints $C = \{(U_1, \tilde{p}_1) \dots (U_k, \tilde{p}_k)\}$, the search objective is to select a subset of constraints $C^* \subseteq C$ to minimize error on L:

$$C^* \leftarrow \operatorname*{argmin}_{C' \subseteq C} E(p_{C'}(y|x), L)$$

where $E(\cdot)$ is a classification error function, and $p_{C'}(y|x)$ is the model fit by label regularization using constraint set C'.

In our experiments, $|C|$ is in the hundreds, so exhaustive, exponential search is impractical. Instead, we consider the following greedy and pseudo-greedy forward-selection algorithms:

- **Greedy (grdy):** Standard greedy search. At each iteration, we select the constraint that leads to the greatest accuracy improvement on L.
- **Semi-greedy (semi):** Rather than selecting the constraint that improves accuracy the most, we randomly select from the top three constraints (Hart and Shogan, 1987).
- **Improved-greedy (imp):** The same as **grdy**, but after each iteration, optionally remove a single constraint. We consider each currently selected constraint, and compute the accuracy attained by removing this constraint from the set. We remove the constraint that improves accuracy the most (if any exists). This constraint is removed from consideration in future iterations.
- **Grasp (grsp):** Greedy Randomized Adaptive Search Procedures (Feo and Resende, 1995) combines **semi** and **imp**.

We run each selection algorithm for 140 iterations (as we discuss below, accuracy plateaus well before then). Then, we select the constraint set that results in the highest accuracy. While this search procedure is computationally expensive, it is fortunately easily parallelizable (by partitioning by constraint), which we take advantage of in our implementation. All constraint selection algorithms use the 40% of the labeled data reserved for training/tuning. After we finalized all models using the tuning data, we then used them to classify the 60% of labeled data reserved for testing.

5 Baselines

We compare label regularization with standard logistic regression (**logistic**) trained using the 40% of labeled data reserved for training/tuning. We also consider several heuristic baselines:

- **Name heuristic, race classification**: We implement the method proposed by (Mohammady and Culotta, 2014), using the top 1000 most popular last names with their race distribution from the U.S. Census Bureau to infer race/ethnicity of users based of most probable race according last name. If the last name is not among the top 1000 most popular for a given race, we simply predict White (the most frequent class).
- **Name heuristic, age classification**: We use the heuristic described in Section 3.3 that estimates a person's age by their first name. Given the age distribution of a first name, we classify the user according to the more probable class.
- **Follower heuristic, political classification**: We reuse the exemplar accounts used in the follower constraint in Section 3.3. That is, rather than using the fact that a user follows "dennis_kucinich" as a soft constraint, we classify such a user as a Democrat. If a user follows more than one of the exemplar accounts, we select the more frequent party.[6] In case of ties (or if the user does not follow any of the accounts), we classify at random.

[6] For the **politician-follower** data the heuristic does not use "thedemocrats" and "gop," because these were used for the original annotation.

	race	age	pol	pol-f	avg
heuristic	43.7	56.0	89.4	65.4	63.6
logistic	81.0	**83.3**	**93.8**	68.7	**81.7**
all-const					
cnt	61.9	45.5	58.1	60.6	56.5
fol	67.3	61.4	**93.8**	60.7	70.8
nam		55.6			
cnt fol	79.4	45.5	79.3	67.9	68.0
cnt nam		44.1			
fol nam		55.9			
cnt fol nam		44.0			
imp-greedy					
cnt	80.1	76.6	65.6	58.9	70.3
fol	76.6	66.1	86.8	69.1	74.7
nam		68.3			
cnt fol	**82.3**	75.2	88.1	**74.3**	80.0
cnt nam		79.2			
fol nam		68.1			
cnt fol nam		75.2			

Table 1: Accuracy on the *testing* set. **all-const** does no constraint selection; **imp-greedy** selects constraints to maximize accuracy on the tuning set using the Improved-greedy algorithm.

Features: For all models, we use a standard bag-of-words representation consisting of a binary term vector for the 200 tweets of each user, their description field, and their name field. We differentiate between terms used in the description, tweet text, and name field, and also indicate hashtags. Finally, we include additional features indicating the accounts followed by each user.

6 Results

Table 1 shows the classification accuracy on the test set for each of the four tasks (F1 results are similar). We begin by comparing **heuristic** and **logistic** to the **all-const** results, which is our proposed label regularization approach using no constraint selection (i.e., no user-labeled data). We can see that for three of the four tasks (**race**, **pol**, **pol-f**), label regularization accuracy is either the same as **logistic** or within 2%. That is, using no user-annotated data, we can obtain accuracy competitive with logistic regression.

For **age**, however, label regularization does quite poorly; only using the **fol** constraints surpasses the heuristic baseline. We suspect that this is in part due to the greater noise in age constraints — Twitter users are particularly non-representative of the overall population according to age. To summarize our answer to **RQ1**, label regularization appears to perform quite well under a moderate amount of constraint noise, but can still fail under excessive noise.

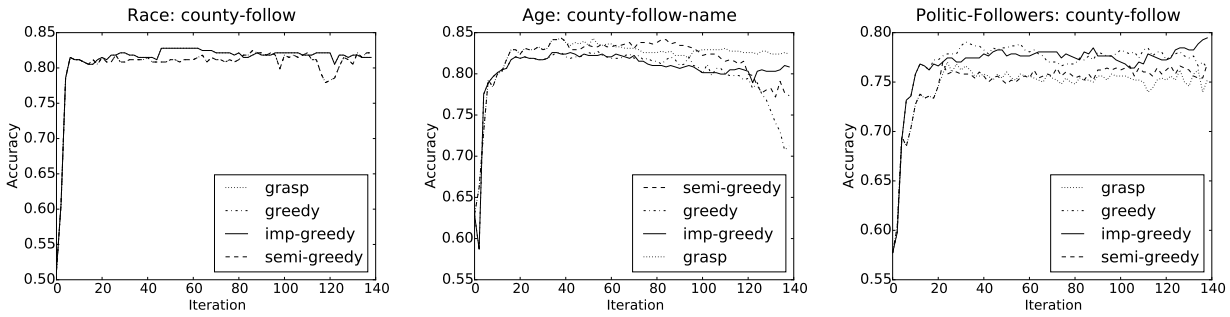
	all	grdy	semi	imp	grsp
Race	77.9	82.5	82.5	**82.8**	**82.8**
Age	48.4	82.8	**84.3**	82.6	**84.3**
Politician	84.0	98.7	96.0	**99.3**	96.7
Politic-fol	61.8	79.1	77.0	**79.5**	77.0
Average	68.0	85.7	85.0	**86.0**	85.2

Table 2: Comparison of the accuracy of constraint selection algorithms on the *tuning* set. **all** uses all possible constraints.

We next consider the effect of the constraint selection algorithms. Table 2 compares the four different constraint selection algorithms, along with the model that selects all constraints. We report the accuracy for each approach considering all constraint types (county, follow, and name, where applicable). Importantly, this accuracy is computed on the *tuning* set, not the test set. The goal here is to determine which search algorithm is able to find the best approximate solution. By comparing with **all**, we can see that constraint selection can significantly improve accuracy on the tuning set (by 18% absolute on average). The differences among the selection algorithms do not appear to be significant.

Figure 1 plots the accuracy at each iteration of constraint select for three of the datasets. The main conclusion we draw from these figures is that high accuracy can be achieved with only a small number of constraints, provided they are carefully chosen. Each method is very close to convergence after using only 20 constraints (selected from hundreds). When examining which constraints are selected, we find that those that apply to many users are often preferred, presumably because there is more data to inform the final model.

Returning to Table 1, we have also listed the accuracy of the **imp-greedy** selection method (which performed best on the tuning set), further strati-



(a) (b) (c)

Figure 1: Accuracy per iteration of constraint selection for three classification tasks.

fied by constraint type. Note that **imp-greedy** selects the constraints that perform best on the tuning set, fits the classification model, and then classifies the testing set. We can see that for three of the four tasks (**race**, **age**, **pol-f**), **imp-greedy** results in higher accuracy than using all the constraints. This is particularly pronounced for **age**: the best result without constraint selection is 61.4, compared with 79.2 for **imp-greedy**. Furthermore, **imp-greedy** outperforms **logistic** on two of four tasks, suggesting that using unlabeled data can improve accuracy. Note that both **imp-greedy** and **logistic** use the same amount of labeled data, though in different ways: **logistic** performs standard supervised classification; **imp-greedy** uses the labeled data to perform constraint selection for label regularization. Thus, to summarize our answer to **RQ2**, we find that **imp-greedy** provides a robust method to select constraints in the presence of noise. While it comes at the cost of a small amount of labeled data, it is less reliant on this data than a traditional supervised approach, and so may be more applicable in streaming settings.

To answer **RQ3**, we can compare the accuracies provided by each of the constraint types in Table 1. For **all-const**, the follower constraints (**fol**) outperform the county constraints (**cnt**) for all tasks, while the name constraint (which only applies to **age**), falls between the two. Including both **cnt** and **fol** improves accuracy on two of the four tasks. These trends change somewhat for **imp-greedy**. The **cnt** constraints are superior for two tasks, while **fol** are superior for the other two. The **nam** constraints again fall between the two. Unlike for **all-const**, using more constraint types improves accuracy on three of four tasks. These differences suggest that the constraint selection algorithms allow label regularization to be more robust to noisy and conflicting constraints. That is, using constraint selection, we can view constraint engineering akin to feature engineering in discriminative, supervised learning methods — developers can add many types of constraints to the model without (much) fear of reducing accuracy. The usual caveat of overfitting applies here as well; indeed, comparing the accuracies on the tuning set (Table 2) with those on the testing set (Table 1) suggests that some over-tuning has occurred, most notably on **age** and **pol**.

We further examined the coefficients of the models trained using each constraint type. We find, for example, that county constraints result in models with large coefficients for location-specific terms (e.g., college names for younger users, southern cities for Republican users), while follower constraints tend to learn models dominated by follower features ("thenation" for Democrats, "glennbeck" for Republicans). Similarly, name constraints result in models dominated by name features. This analysis helps explain how combining constraint types can improve overall accuracy, since each type emphasizes different subsets of features.

This difference between constraint types is further shown in Table 3, which lists the top features for the semi-greedy constraint selection algorithm, fit using different subsets of constraints. In this table, the italicized words are the words from the description field of the user's profile, the underlined words are followed accounts, and the bold words are the words

age	**under 25**	**above 25**
County	athens *tech uga virginia georgia*	airport nashvillescene at theonion *and*
Follow	altpress colourlovers hotnewhiphop planetminecraft me	newsobserver baseballamerica peopleenespanol breakingnews hogshaven
Name	**katherine diana** me my this	**debra lori sandra janet** *No Desc*
politician	**Democratic**	**Republican**
County	*oregon* eugene oregon nesn *university*	colts beach tahoe *indiana* jgfortwayne
Follow	keithellison repjohnlewis sensanders thinkprogres thenation	gopleader senmikelee senrandpaul gop glennbeck

Table 3: Top features learned by label regularization for the age and politician datasets using semi-greedy constraint selection. Models were fit separately for each constraint type (county, follow, name). Italicized words are from the description field, bold words are from the name field, and underlined words are followed accounts.

from the name field of the user profile. In the first row, we display the top features for a model fit using only county constraints. College names appear as top features for younger users, and "airport" and @NashvilleScene (a newspaper) are for older users. The second row of Table 3 shows the top features for following constraints; some news channels are appear for younger (Alternative Press) and older (The News & Observer) users. The third row shows the top features for name constraints, and some names are in the top features for younger (Katherine and Diana) and older (Debra, Lori, Sandra, and Janet). In addition, the absence of a profile description is indicative of older users.

The bottom of Table 3 shows top features for the politician dataset. The first row shows that some colleges, a sports network in New England, and locations in the Pacific Northwest are indicative of Democrats. Indiana-related terms are strong indicators of Republicans: *indiana*, the Indianapolis Colts (an American football team), and 'jgfortwayne' (The Journal Gazette, a newspaper in Fort Wayne, Indiana). This aligns with the strong support of the Republican party in Indiana.[7] The second row shows top-ranked following features. Accounts 'keithellison' and 'repjohnlewis' are top features for Democratic Party; these belong to Keith Ellison and John Robert Lewis, members of the Democratic leadership of the House of Representatives. On other hand, the 'gopleader' (the official account for the Republican's majority leader in the House) and 'senmikelee' (Republican Senator Mike Lee from Utah) are the top features for Republicans.

7 Conclusions and Future work

While label regularization has been used on a number of NLP tasks, we have presented evidence that it is applicable to latent attribute inference even using many noisy, heterogeneous constraints. We have compared a number of constraint selection algorithms and found they can make label regularization more robust to noisy constraints, allowing developers to combine many rich constraint types without reducing accuracy.

There are many avenues for future work. Most pressing is the need to directly address the sampling bias created when constraints derived from the overall population are applied to online users. We plan to explore alternative optimization strategies to explicitly address this issue. Finally, additional research should quantify how responsive label regularization approaches are to the changing linguistic patterns common in online data.

References

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

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

[7] http://en.wikipedia.org/wiki/Politics_of_Indiana

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.

M. Chang, L. Ratinov, and D. Roth. 2007. Guiding semi-supervision with constraint-driven learning. In *ACL*, pages 280–287, Prague, Czech Republic, 6. Association for Computational Linguistics.

Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. epluribus: Ethnicity on social networks. In *ICWSM*.

Ming-Wei Chang, Lev Ratinov, and Dan Roth. 2012. Structured learning with constrained conditional models. *Machine learning*, 88(3):399–431.

Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on twitter: It's not easy! In *ICWSM*.

Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. 2011. Predicting the political alignment of twitter users. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 192–199. IEEE.

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

Gregory Druck, Gideon Mann, and Andrew McCallum. 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*, pages 595–602.

Gregory Druck, Gideon Mann, and Andrew McCallum. 2009. Semi-supervised learning of dependency parsers using generalized expectation criteria. In *ACL*.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 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, page 13651374, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas A Feo and Mauricio GC Resende. 1995. Greedy randomized adaptive search procedures. *Journal of global optimization*, 6(2):109–133.

Kuzman Ganchev and Dipanjan Das. 2013. Cross-lingual discriminative learning of sequence models with posterior regularization. In *EMNLP*, pages 1996–2006.

Kuzman Ganchev, Jennifer Gillenwater, and Ben Taskar. 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*, pages 369–377. Association for Computational Linguistics.

Kuzman Ganchev, Joo Graca, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11:20012049, August.

Andrew Gelman. 2007. Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

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

Joao Graça, Kuzman Ganchev, and Ben Taskar. 2007. Expectation maximization and posterior constraints. In *NIPS*, volume 20, pages 569–576.

Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182.

Aria Haghighi and Dan Klein. 2006. Prototype-driven learning for sequence models. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 320–327. Association for Computational Linguistics.

J Pirie Hart and Andrew W Shogan. 1987. Semi-greedy heuristics: An empirical study. *Operations Research Letters*, 6(3):107–114.

Rong Jin and Yi Liu. 2005. A framework for incorporating class priors into discriminative classification. In *In PAKDD*.

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

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*, pages 1110–1119.

Amanda Lenhart and Susannah Fox. 2009. Twitter and status updating. pew internet & american life project.

Percy Liang, Michael I. Jordan, and Dan Klein. 2009. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 641648, New York, NY, USA. ACM.

Songrit Maneewongvatana and David M Mount. 2002. Analysis of approximate nearest neighbor searching with clustered point sets. *Data Structures, Near Neighbor Searches, and Methodology*, 59:105–123.

Gideon S. Mann and Andrew McCallum. 2007. Simple, robust, scalable semi-supervised learning via expectation regularization. In *Proceedings of the 24th Inter-*

national Conference on Machine Learning, ICML '07, page 593600, New York, NY, USA. ACM.

Gideon S. Mann and Andrew McCallum. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *J. Mach. Learn. Res.*, 11:955984, March.

Prem Melville, Wojciech Gryc, and Richard D. Lawrence. 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, page 12751284, New York, NY, USA. ACM.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, Barcelona, Spain.

Ehsan Mohammady and Aron Culotta. 2014. Using county demographics to infer attributes of twitter users. In *ACL Joint Workshop on Social Dynamics and Personal Attributes in Social Media*.

D.R. Musicant, J.M. Christensen, and J.F. Olson. 2007. Supervised learning by training on aggregate outputs. In *Seventh IEEE International Conference on Data Mining, 2007. ICDM 2007*, pages 252–261.

Dong Nguyen, Noah A. Smith, and Carolyn P. Ros. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, page 115123, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to polls: Linking text sentiment to public opinion time series. In *International AAAI Conference on Weblogs and Social Media*, Washington, D.C.

Huseyin Oktay, Aykut Firat, and Zeynep Ertem. 2014. Demographic breakdown of twitter users: An analysis based on names. In *Academy of Science and Engineering (ASE)*.

Giorgio Patrini, Richard Nock, Tiberio Caetano, and Paul Rivera. 2014. (almost) no label no cry. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 190–198. Curran Associates, Inc.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, and Quoc V. Le. 2009. Estimating labels from label proportions. *J. Mach. Learn. Res.*, 10:23492374, December.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*, SMUC '10, page 3744, New York, NY, USA. ACM.

Delip Rao, Michael J. Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press.

Robert E. Schapire, Marie Rochery, Mazin G. Rahim, and Narendra K. Gupta. 2002. Incorporating prior knowledge into boosting. In *Proceedings of the Nineteenth International Conference*, pages 538–545.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E P Seligman, and Lyle H Ungar. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PloS one*, 8(9):e73791. PMID: 24086296.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478. Association for Computational Linguistics.

Nate Silver and Allison McCanc. 2014. How to tell someone's age when all you know is her name. Retrieved from http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name/.

Mengqiu Wang and Christopher D. Manning. 2014. Cross-lingual projected expectation regularization for weakly supervised learning. *TACL*, 2:55–66.

Zuoguan Wang, Siwei Lyu, Gerwin Schalk, and Qiang Ji. 2012. Learning with target prior. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2231–2239. Curran Associates, Inc.

Samuel Craig Watkins. 2009. *The young and the digital: what the migration to social-network sites, games, and anytime, anywhere media means for our future*. Beacon Press.

Jun Zhu, Ning Chen, and Eric P Xing. 2014. Bayesian inference with posterior regularization and applications to infinite latent svms. *Journal of Machine Learning Research*, 15:1799–1847.