# Collective Supervision of Topic Models for Predicting Surveys with Social Media

**Adrian Benton**
Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218
adrian@cs.jhu.edu

**Michael J. Paul**
College of Media, Communication, and Information
University of Colorado, Boulder, CO 80309
mpaul@colorado.edu

**Braden Hancock**
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
braden.hancock@stanford.edu

**Mark Dredze**
Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD 21211
mdredze@cs.jhu.edu

## Abstract

This paper considers survey prediction from social media. We use topic models to correlate social media messages with survey outcomes and to provide an interpretable representation of the data. Rather than rely on fully unsupervised topic models, we use existing aggregated survey data to inform the inferred topics, a class of topic model supervision referred to as *collective* supervision. We introduce and explore a variety of topic model variants and provide an empirical analysis, with conclusions of the most effective models for this task.

## Introduction

Social media has proved invaluable for research in social and health sciences, including sociolinguistics (Eisenstein, Smith, and Xing 2011), political science (O'Connor et al. 2010), and public health (Paul and Dredze 2011). A common theme is the use of topic models (Blei, Ng, and Jordan 2003), which, by identifying major themes in a corpus, summarize the content of large text collections. Topic models have been applied to characterize tweets (Ramage, Dumais, and Liebling 2010), blog posts and comments (Yano, Cohen, and Smith 2009; Paul and Girju 2009), and other short texts (Phan, Nguyen, and Horiguchi 2008).

One goal of social media analytics is to complement or replace traditional survey mechanisms (Thacker and Berkelman 1988; Krosnick, Judd, and Wittenbrink 2005). Traditional phone surveys are both slow and expensive to run. For example, the CDC's annual Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that collects health data by calling more than 400,000 Americans. The survey costs millions of dollars to run each year, so adding new questions or obtaining finer-grained temporal information can be prohibitive. Expanding this survey through social media monitoring promises methods that are both fast and cheap. For example, research has shown that social media features can complement existing models in predicting health statistics at the collective, county level (Culotta 2014). Myslín et al. (2013) have also shown that Twitter can be used to monitor opinions on tobacco use.

Paul and Dredze (2011) showed initial work on correlating several BRFSS questions with social media topics learned by a topic model. However, since topic models are an unsupervised method for learning text representations, they may not identify topics that naturally correspond to those useful in a survey. One solution is to provide supervision to the models based on previously conducted, or topically related, surveys. This can include using surveys of existing US states to train topic models for predicting missing US states, or using previous surveys on similar issues when conducting a new survey.

Numerous topic models incorporate supervision, such as predicting labels for each document, e.g., supervised LDA (Mcauliffe and Blei 2008); modeling tags associated with each document, e.g., labeled LDA (Ramage et al. 2009) or tagLDA (Zhu, Blei, and Lafferty 2006); placing priors over topic-word distributions (Jagarlamudi, Daumé III, and Udupa 2012; Paul and Dredze 2013); or interactive feedback from the user (Hu et al. 2014). However, none of these models support the aggregate-level labels provided by surveys.

We present a collective method of supervision for topic models, where aggregate-level labels are provided for groups of messages instead of individual messages. We experiment with a variety of modifications to topic models to support this task. We evaluate our methods on using Twitter to predict three survey questions taken from the annual BRFSS survey. We show that incorporating aggregate data leads to more predictive topics.

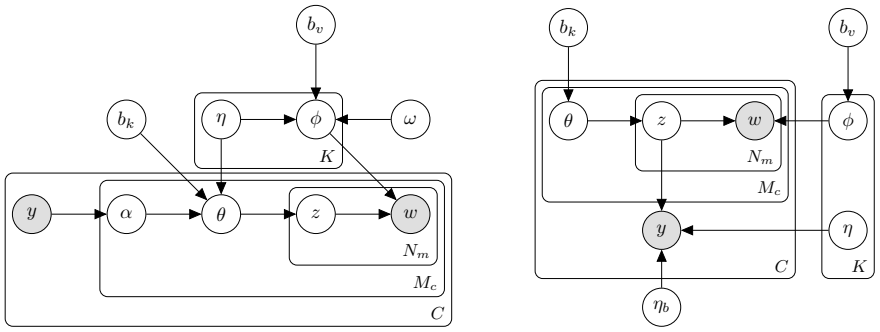
## Collective Supervision

We define **collective** supervision[1] as supervision in which labels are provided for *groups* or *collections* of documents, rather than supervision at the level of individual documents. Example collections include particular geographic areas (e.g., U.S. states) or time periods (e.g., weeks). Examples of collective supervision include the proportion of smokers or the number of gun owners in each U.S. state. While our formulation is general, we focus on geographic areas as collections, taking supervision from U.S. surveys.

Under this framework, a corpus is partitioned into $C$ collections, where the $j$th collection is labeled with a value, $y_j$ (e.g., the percentage of smokers in location $j$). The $m$th document is associated with a collection index, $c_m$.



[1] We borrow the name for this type of supervision from *collective graphical models* (Sheldon and Dietterich 2011).




(a) Upstream Model (b) Downstream Model

Figure 1: Graphical model of (a) the most complex upstream model, with adaptive supervision and structured word distributions, and (b) the downstream collective sLDA model. Constants determining strength of prior on $\alpha$, $b_k$, $b_v$, $\eta$, and $\omega$ are omitted due to space constraints.

## Models

This section presents topic models based on Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). We follow standard LDA notation in this section. LDA is a fully unsupervised model, which may have limited utility when trying to learn topics that are predictive of surveys. This paper will show how to incorporate collective supervision into topic models. This section presents several models that incorporate collective supervision in different ways. We classify these models as either **upstream** or **downstream**, using the terminology of Mimno and McCallum (2008), referring to whether the response variables $\boldsymbol{y}$ are generated before or after the text in the generative stories.

Both types of models contain survey coefficients for each topic, $\eta_k$, indicating whether a topic is positively or negatively correlated with the survey values. In upstream models, these coefficients inform the priors over topic distributions in documents, while in downstream models, these are used to predict the survey values directly.

We emphasize that our focus is not on creating novel topic models, but on an evaluation of the most effective model formulation for the proposed task. Therefore, we consider modifications to different types of existing topic models.

### Upstream Models

In upstream topic models, supervision influences the *priors* over topic distributions in documents. This is done with Dirichlet-multinomial regression (DMR) (Mimno and McCallum 2008), in which the Dirichlet parameters are log-linear functions of the document labels $\boldsymbol{y}$ and the regression coefficients $\boldsymbol{\eta}$. Under a DMR topic model (the **upstream** model in our experiments), each document has its own Dirichlet($\tilde{\theta}_m$) prior, with $\tilde{\theta}_{mk}$=$\exp(b_k + y_m\eta_k)$, where $y_m$ is the label of the $m$th document, $\eta_k$ is the $k$th topic's survey coefficient, and $b_k$ is a bias term (intercept). For positive $\eta_k$, the prior for topic $k$ in document $m$ will increase as $y_m$ increases, while negative $\eta_k$ will decrease the prior.

We straightforwardly adapt this DMR model to the collective setting by replacing document labels with collection labels: $y_m = y_{c_m}$. In this version, the prior over topics is informed by the *collection* of the document. Figure 2 provides the generative story, including variants below.

1. For each document $m$:
   (a) $y_{c_m}$ is the feature value associated with the document's collection $c_m$
   (b) $\alpha_m \sim \mathcal{N}(y_{c_m}, \sigma_\alpha^2)$ (**adaptive** version)
   or
   $\alpha_m = y_{c_m}$ (standard version)
   (c) $\tilde{\theta}_{mk} = \exp(b_k + \alpha_m\eta_k)$, for each topic $k$
   (d) $\theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
2. For each topic $k$:
   (a) $\tilde{\phi}_{kv} = \exp(b_v + \omega_v\eta_k)$, for each word $v$ (**words** version)
   or
   $\tilde{\phi}_{kv} = \exp(b_v)$ (standard version)
   (b) $\phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$
3. For each token $n$ in each document $m$:
   (a) Sample topic index $z_{mn} \sim \theta_m$
   (b) Sample word token $w_{mn} \sim \phi_{z_{mn}}$

Figure 2: Generative story for the various upstream models.

**Adaptive Supervision** A limitation of the basic upstream model is the assumption that the $\boldsymbol{y}$ values are always available and accurate. This is often not true; for example, a large percentage of social media messages cannot be resolved to specific locations, so their collections are unknown.

We thus experiment with a novel variant of the standard DMR model that allows the response variables to deviate from the input values. We replace each document's supervised variable with an auxiliary variable $\alpha_m$ that is normally distributed with mean $y_{m_c}$, as described in step 1b of the generative story. This encourages the value to be near the input, while adapting to the data. Models with adaptive variables are denoted with **ada**.

**Structured Word Distributions** We also experiment with variants in which the collective supervision indirectly influences each topic's distribution over words, in addition to each document's distribution over topics. For example, we might *a priori* believe that topics associated with high gun ownership are more likely to have words in common.

We do this using the structured-prior topic modeling framework of SPRITE (Paul and Dredze 2015), which extends DMR to use log-linear priors in various ways. In this

1. For each document $m$:
   (a) $\tilde{\theta}_{mk} = \exp(b_k)$, for each topic $k$
   (b) $\theta_m \sim \text{Dirichlet}(\tilde{\theta}_m)$
2. For each topic $k$:
   (a) $\tilde{\phi}_{kv} = \exp(b_v)$
   (b) $\phi_k \sim \text{Dirichlet}(\tilde{\phi}_k)$
3. For each token $n$ in each document $m$:
   (a) Sample topic index $z_{mn} \sim \theta_m$
   (b) Sample word token $w_{mn} \sim \phi_{z_{mn}}$
4. For each document collection $j$:
   (a) Let $\overline{z}_{jk}$ be the average proportion of topic $k$ in collection $j$
   (b) $y_j \sim \mathcal{N}(\eta_b + \eta^T \overline{z}_j, \sigma_y^2)$

Figure 3: Generative story for the collective sLDA model. sLDA is a special case where each document belongs to a unique collection.

model, the topic survey coefficients $\boldsymbol{\eta}$ are used in the priors over word distributions in addition to topic distributions, as described in step 2a of the generative story. Words with a positive value of $\omega_v$ will have a higher prior in topics with a positive value of $\eta_k$. These model variants are denoted with **words** in the name.

## Downstream Models

Downstream topic models generate the response variables $\boldsymbol{y}$ after the text, conditioned on the topic values. Supervised LDA (**sLDA**) (Mcauliffe and Blei 2008) falls in this category. sLDA follows the same generative story as LDA for generating a corpus, and generates each document's response variable $y_m$ as the output of a linear model conditioned on the document's average topic counts: $y_m \sim \mathcal{N}(\eta_b + \eta^T \overline{z}_m, \sigma_y^2)$, where $\eta_b$ is a bias term. We experiment with adapting sLDA to the collective setting by setting $y_m = y_{c_m}$, pretending that each document has an observed value corresponding to its collection's value.

We also experiment with a novel **collective** variant of sLDA that correctly models the response at the collection level. Each collection $j$ has a single response variable (rather than one variable per document) that depends on the average topic counts across the entire collection: $y_j \sim \mathcal{N}(\eta_b + \eta^T \overline{z}_j, \sigma_y^2)$, where $\overline{z}_j$ is the average count of topic assignments in all $j$th collection documents. The collective variant is equivalent to sLDA if each document has a unique collection index. Figure 3 provides the generative story.

## Parameter Estimation

Inference for each model involves alternating between one iteration of Gibbs sampling (sampling each token's topic assignment) and one iteration of gradient ascent for the parameters $b$, $\alpha$, $\omega$, and $\eta$. We include 0-mean Gaussian priors on all parameters to prevent overfitting. See Paul and Dredze (2015) for the upstream gradient updates.

The downstream Gibbs samplers augment each token's sampling distribution with the likelihood of the response, $\mathcal{N}(\eta_b + \eta^T \overline{z}_{c_m}, \sigma_y^2)$, where the counts $\overline{z}_{c_m}$ include the topic being considered. The variance $\sigma_y^2$ controls how strongly the response likelihood affects the choice of topic.

| Dataset | Vocab | State | County | BRFSS |
|---|---|---|---|---|
| Guns | 12,358 | 29.7% | 18.6% | Owns firearm |
| Vaccines | 13,451 | 23.6% | 16.2% | Had flu shot |
| Smoking | 13,394 | 19.6% | 12.8% | Current smoker |

Table 1: A summary of the three datasets: size of the vocabulary, proportion of messages tagged at the state and county level, and the state-level survey question (BRFSS) asked.

# Experiments

## Data

We evaluate the prediction of three survey questions using Twitter data. The survey questions are from BRFSS, an annual phone survey of hundreds of thousands of American adults, chosen for its very large and geographically widespread sample. We selected the following three questions: the percentage of respondents in each U.S. state who (1) have a firearm in their house (data from 2001, when the question was last asked), (2) have had a flu shot in the past year (from 2013), and (3) are current smokers (from 2013). Our goal is to predict state-level results (guns, vaccinations, smoking) based on topic representations of Twitter data.

We created three Twitter datasets based on keyword filtering with data collected from Dec. 2012 through Jan. 2015 to match tweets relevant to these three survey questions. We selected 100,000 tweets uniformly at random for each dataset and geolocated them to state/county using *Carmen* (Dredze et al. 2013). Geolocation coverage is shown in Table 1. We experimented with two sources of collective supervision:

**Survey** The direct type of collective supervision is to use the values of the BRFSS survey responses that we are trying to predict. Each U.S. state is a collection, and each collection includes tweets resolved to that state. This setting reflects predicting the values for some states using data already available from other states. This setting is especially relevant for BRFSS, since the survey is run by each state with results collected and aggregated nationally. Since not all states run their surveys at the same time, BRFSS routinely has results available for some states but not yet others.

**Census** We also experimented with an alternative, *indirect* type of collective supervision, in the form of demographic information from the 2010 U.S. Census. Demographic variables are correlated with the responses to the surveys we are trying to predict (Hepburn et al. 2007; King, Dube, and Tynan 2012; Gust et al. 2008), so we hypothesize that using demographic information may lead to more predictive and interpretable topic models than no supervision at all. This approach may be advantageous when domain-specific survey information is not readily available.

From the Census, we used the percentage of white residents per county, for tweets whose county could be resolved. Although this feature is not directly related to our dependent variable, it is sampled at a finer granularity than the state-level survey feature. Proportion of tweets tagged with this feature are also included in Table 1. In our experiments we consider these two types of supervision in isolation to assess the usefulness of each class of distant supervision.

| Features | Model | Guns | | Vaccines | | Smoking | |
|---|---|---|---|---|---|---|---|
| None | LDA | 17.44 | 2313 ($\pm$52) | 8.67 | 2524 ($\pm$20) | 4.50 | 2118 ($\pm$5) |
| Survey | Upstream | 15.37 | 1529 ($\pm$12) | 6.54 | 1552 ($\pm$11) | 3.41 | 1375 ($\pm$6) |
| | Upstream-words | 11.50 | **1429** ($\pm$22) | 6.37 | 1511 ($\pm$57) | 3.41 | 1374 ($\pm$2) |
| | Upstream-ada | 11.48 | 1506 ($\pm$67) | 5.82 | **1493** ($\pm$49) | 3.41 | 1348 ($\pm$6) |
| | Upstream-ada-words | **11.47** | 1535 ($\pm$28) | 7.20 | 1577 ($\pm$15) | **3.40** | 1375 ($\pm$3) |
| | Downstream-sLDA | 11.52 | 1561 ($\pm$22) | 11.22 | 1684 ($\pm$7) | 3.95 | 1412 ($\pm$3) |
| | Downstream-collective | 12.81 | 1573 ($\pm$20) | 9.17 | 1684 ($\pm$6) | 4.35 | 1412 ($\pm$4) |
| Census | Upstream | 11.51 | 1555 ($\pm$27) | **5.15** | 1575 ($\pm$90) | 3.42 | 1377 ($\pm$8) |
| | Upstream-words | 15.88 | 1440 ($\pm$38) | 6.85 | 1549 ($\pm$57) | 3.41 | 1376 ($\pm$5) |
| | Upstream-ada | 11.50 | 1534 ($\pm$48) | 6.49 | 1509 ($\pm$21) | 3.41 | **1346** ($\pm$7) |
| | Upstream-ada-words | 11.50 | 1553 ($\pm$20) | 6.35 | 1584 ($\pm$19) | 3.41 | 1378 ($\pm$3) |
| | Downstream-sLDA | 11.52 | 1586 ($\pm$20) | 8.83 | 1688 ($\pm$7) | 5.37 | 1411 ($\pm$3) |
| | Downstream-collective | 15.61 | 1586 ($\pm$44) | 9.15 | 1681 ($\pm$10) | 4.72 | 1412 ($\pm$3) |

Table 2: RMSE of the prediction task (left) and average perplexity (right) of topic models over each dataset, $\pm$ stddev. Perplexity is averaged over 5 sampling runs and RMSE is averaged over 5 folds of U.S. states. For comparison, the RMSE on the prediction task using a bag-of-words model was 11.50, 6.33, and 3.53 on the Guns, Vaccine, and Smoking data, respectively.

| **Model Class 1** | **Model Class 2** | **Prediction (MSE)** | | **Perplexity** | |
|---|---|---|---|---|---|
| Downstream | Upstream | 466 (0.001) | 2.54 (0.014) | 75 (0.000) | 5.59 (0.000) |
| Census | Survey | 1516 (0.032) | -0.15 (0.882) | 1266 (0.003) | 1.07 (0.287) |
| Direct supervision | Adaptive supervision | 879 (0.791) | 0.08 (0.938) | 760 (0.345) | -3.48 (0.001) |
| Upstream with **words** | Upstream without **words** | 810 (0.440) | -1.62 (0.110) | 833 (0.695) | -3.02 (0.004) |
| Downstream-sLDA | Downstream-collective | 133 (0.041) | 2.21 (0.035) | 228 (.923) | -.240 (.812) |

Table 3: Performance comparison for different model/feature classes. The first set of numbers in each cell is the Wilcoxon signed-rank statistic and corresponding p-value. The second set is the paired t-test statistic and corresponding p-value. A positive sign of the t-test statistic indicates that Model Class 1 has higher prediction error or perplexity than Model Class 2.

## Experimental Details

We tuned each model for held-out perplexity and evaluated its ability to predict the survey proportion for each state. Held-out perplexity was computed using the "document completion" approach (Wallach et al. 2009); specifically, every other token was used for training, with perplexity measured on the remaining tokens. We also compared to an LDA model (no supervision).

For tuning, we held out 10,000 tweets from the guns dataset and used the best parameters for all datasets. We ran Spearmint (Snoek, Larochelle, and Adams 2012) for 100 iterations to tune the learning parameters, running each sampler for 500 iterations. Spearmint was used to tune the following learning parameters: the initial value for $b$, and the variance of the Gaussian regularization on $b$, $\boldsymbol{\eta}$, $\boldsymbol{\omega}$, $\boldsymbol{\alpha}$, and $\boldsymbol{y}$ (in the downstream model). Once tuned, all models were trained for 2000 iterations, using AdaGrad (Duchi, Hazan, and Singer 2011) with a master step size of 0.02.

We evaluated the utility of topics as features for predicting the collective survey value for each U.S. state, reflecting how well topics capture themes relevant to the survey question. We inferred $\theta_m$ for each tweet and then averaged these topic vectors over all tweets originating from each state, to construct 50 feature vectors per model. We used these features in a regularized linear regression model. Average root mean-squared error (RMSE) was computed using five-fold cross-validation: 80% of the 50 U.S. states were used to train, 10% to tune the $\ell_2$ regularization coefficient, and 10% were used for evaluation. In each fold, the topic models used supervision only for tweets from the training set states, while the $\boldsymbol{y}$ values were set to 0 (a neutral value) for the held-out states.

We swept over the $\ell_2$ regularization coefficient. For both perplexity and prediction performance, we sweep over number of topics in $\{10, 25, 50, 100\}$ and report the best result. Results are averaged across five sampling runs. For supervised models, we use either the survey value or Census demographic value as supervision.

The text was preprocessed by removing stop words and low-frequency words. We applied z-score normalization to the BRFSS/Census values within each dataset, so that the mean value was 0. For tweets whose location could not be resolved, the value was set to 0 for the upstream models. In the downstream models, such tweets are assigned to a dummy collection whose response likelihood is fixed to 1.

## Results

Results are shown in Table 2. The important takeaway is that topic models with collective supervision are more predictive than LDA, an unsupervised model. Not only do the supervised models substantially reduce prediction error, as might be expected, but they also have substantially lower perplexity, and thus seem to be learning more meaningful concepts.

The poor performance of LDA may be partially explained by the fact that Spearmint seems to overfit LDA to the tuning set. Other models attained a tuning set perplexity of between 1500 to 1600, whereas LDA attained 1200. To investigate this issue further, we separately ran experiments with hand-tuned models, which gave us better held-out results for LDA, though still worse than the supervised topic models (e.g., RMSE of 16.44 on the guns data). Although Spearmint

| Guns | | Vaccines | | Smoking | |
|---|---|---|---|---|---|
| $r = -1.04$ | $r = 0.43$ | $r = -0.25$ | $r = 1.07$ | $r = -0.62$ | $r = 1.04$ |
| gun | guns | ebola | truth | smoking | #cigar |
| mass | people | trial | autism | quit | #nowsmoking |
| shootings | human | vaccines | outbreak | stop | #cigars |
| call | get | promising | science | smokers | cigar |
| laws | would | experimental | know | #quitsmoking | james |
| democrats | take | early | connection | best | new |
| years | one | first | via | new | thank |
| since | away | results | knows | help | beautiful |

Table 4: Sample topics for the *Upstream-ada-words* model supervised with the survey feature. A topic with a strongly negative as well as a strongly positive $r$ value was chosen for each dataset.

tuning is not perfect, it is fair to all models.

For additional comparison, we experimented with a standard bag-of-words model, where features were normalized counts across tweets from each state. This comparison is done to contextualize the magnitude of differences between models, even though our primary goal is to compare different types of topic models. We found that the bag-of-words results (provided in the caption of Table 2) are competitive with the best topic model results. However, topic models are often used for other advantages, e.g., interpretable models.

**Comparing Model Variants** We now compare the different variants of the collectively supervised models. We measured the significance of the differences in performance of different classes of models according to (i) a Wilcoxon signed-rank test, and (ii) a paired t-test. Model results were paired within each dataset, fold (for the prediction task, since mean squared error varied from fold-to-fold), and model class or feature set if applicable. For example, to compare the model variants with adaptive supervision to those with direct supervision, we paired the results from *Upstream* with *Upstream-ada*, and *Upstream-words* with *Upstream-ada-words*. There was not a one-to-one correspondence between upstream models and downstream models, so to compare these two classes, we paired each downstream variant with a randomly selected upstream variant. The test statistics and p-values are shown in Table 3.

Surprisingly, the upstream models performed better than downstream in nearly every case, even though the downstream models are directly trained to predict the response. (This was true with multiple sLDA implementations.) The differences between the two types of models are highly significant under all tests and metrics.

Comparing the two downstream models, we find that the results are mixed, but after pairing the results across folds, the significance tests indicate that the collective downstream model has significantly lower prediction error (with $p < .05$ under both tests) than sLDA, although the perplexity results are statistically indistinguishable.

Comparing the upstream models, the differences between variants with direct/adaptive supervision and variants with/without words are only significant for perplexity and only under a t-test. Thus, these variants may offer improved representations of the text, but the differences are minor.

Comparing models trained with Survey versus Census data, both improve over LDA and obtain similar results, with the Survey models generally performing better: the prediction and perplexity results are both significantly different, but only under the Wilcoxon signed-rank test. The strong results using only Census data suggest that our methods can still have utility in scenarios when there is limited survey data available, but for which the survey questions have demographic correlates.

**Qualitative Inspection** Table 4 displays example topics learned by the richest model: *Upstream-ada-words*. For example, a topic about the results of the ebola vaccine trials is negatively correlated with vaccine refusal, while a topic about the connection between vaccines and autism is positively correlated with vaccine refusal. We did not observe noticeable qualitative differences in topics learned by the different models, with an exception of LDA, where the topics tended to contain more general words and fewer hashtags than topics learned by the supervised models.

## Use Case: Predicting Support for Gun Restrictions

We ran a final experiment to consider the setting of predicting a new survey with limited available data. We chose the subject of requiring universal background checks for firearm purchases, a topic of intense interest in the U.S. in 2013 due to political events. Despite the national interest in this topic, telephone surveys were only conducted for less than half of U.S. states. We identified 22 individual state polls in 2013 that determined the proportion of respondents that opposed universal background checks. 15 of the states were polled by Public Policy Polling, while the remaining 7 states were polled by Bellwether Research, Nelson A. Rockefeller Research, DHM Research, Nielsen Brothers, Repass & Partners, or Quinnipiac University. We take this as a real-world example of our intended setting: a topic of interest where resources limited the availability of surveys.

We used a topic model trained with data from the universal background check (UBC) survey question as features for predicting the state values for the UBC surveys. For this experiment, we focused on the best-performing topic model from the previous section: *Upstream-ada-words*. As in the previous experiments, we used topic features in a linear regression model, sweeping over $\ell_2$ regularization constants and number of topics, and we report test performance of the best-performing settings on the tuning set. We evaluated the model using five-fold cross-validation on the 22 states.

Additionally, we sought to utilize data from a previous,

| Features | Model | RMSE (2001 Y included) | RMSE (2001 Y omitted) |
|---|---|---|---|
| None | No model | 7.26 | 7.59 |
| | Bag of words | 5.16 | 7.31 |
| | LDA | 6.40 | 7.59 |
| Survey | Upstream-ada-words | **5.11** | **5.48** |

Table 5: RMSE when predicting proportion respondents opposing universal background checks with topic distribution features. We experimented with (left) and without (right) including the 2001 proportion households with a firearm survey data as an additional feature. "*No model*" is the regression where we predict using only the 2001 proportion of households with a firearm.

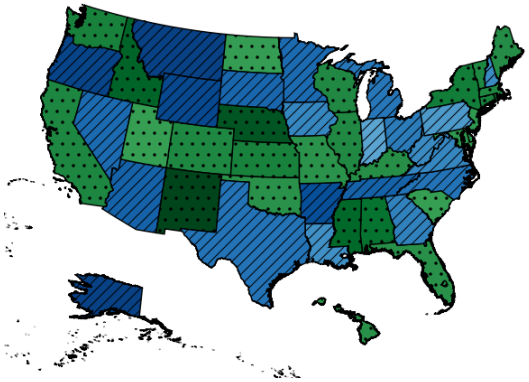


Figure 4: Predictions from the *Upstream-ada-words* model trained on the proportion opposed to universal background checks. The 22 blue states hatched with lines were in the model's training set, while we have no survey data for the 28 green, dotted states. Darker colors denote higher opposition to background checks.

topically-related survey: the "Guns" BRFSS survey used in the previous section, which measured the proportion of households with a firearm, asked in 2001. While the survey asks a different question, and is several years out of date, our hypothesis is that the results from the 2001 survey will be correlated with the new survey, and thus will be a good predictor. We experimented with and without including the values of the 2001 BRFSS survey (which is available for all 50 states) as an additional feature in the regression model.

Table 5 contains the cross-validation test results. We compared the supervised topic model performance to LDA as well as a bag-of-words model. To put the results in context, we also trained regression models using only the 2001 BRFSS values as features ("No model, 2001 Y included") as well as a regression model with no features at all, only an intercept ("No model, 2001 Y omitted").

In general, models that use text features outperform the baseline using only data from the 2001 survey, showing that text information derived from social media can improve survey estimation, even when using topically-related historic data. Moreover, the collectively supervised *Upstream-ada-words* model trained on the UBC survey data is significantly better than an unsupervised topic model (LDA) with $p = 0.06$, under a paired t-test across folds. The difference between *Upstream-ada-words* and the bag-of-words model is not significant ($p = 0.16$), although the difference is larger in the setting where the 2001 survey data is omitted. For the Public Policy Polling surveys used to build the UBC data, the margin of error ranged from 2.9% (more than 1000 polled) to 4.4% (500 polled). An RMSE of 5.1 is approximately equivalent to a 10% margin of error at the 95% confidence level, similar to polling roughly 100 people.

We also investigated the utility of using topic models trained on the topically-related 2001 BRFSS firearm data, rather than the target 2013 UBC data. The potential advantage is that the 2001 data is available for all 50 states, so more data is available to train the topic models. However, training the topic model on this data resulted in worse RMSEs: 6.66 and 6.60, compared to 5.11 and 5.48. Thus, in this case, it was more effective to train the topic models on the target data, even though less data was available.

Finally, we trained our *Upstream-ada-words* regression model (with 2001 BRFSS features) on all 22 states, and used this model to make predictions of opposition to universal background checks for the remaining 28 states. The predictions are shown in Figure 4.[2]

## Discussion and Conclusion

We have presented a wide range of topic models to predict aggregate-level survey scores from messages in social media, and we have shown how topic models can be used or modified for this setting. Our results offer guidance for topic modeling in collective settings:

- Upstream models had substantially better prediction error and perplexity than downstream models, and these differences are highly significant under both types of significance tests. We therefore recommend using upstream topic models for modeling surveys.
- The collective variant of sLDA had lower prediction error than the standard sLDA model by a statistically significant amount, though this varied by dataset, and is still worse than the upstream models. The perplexity of the two variants was about the same.
- We found weak evidence that our modifications to the basic upstream model—using adaptive supervision and structured word priors—offer some advantages, though the improvements were not significant in most cases.

In addition to conducting experiments to compare different models, we also applied our best-performing model to a real-world task: predicting public opinion on gun restrictions (in the form of universal background checks) in U.S. states where phone surveys have not been asked with this question. We thus offer this approach as a general methodology for using social media to estimate resource-limited surveys.

[2]Exact values, as well as our datasets can be found at `https://github.com/abenton/collsuptmdata`

## References


Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*.

Culotta, A. 2014. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1335–1344. ACM.

Dredze, M.; Paul, M.; Bergsma, S.; and Tran, H. 2013. Carmen: A Twitter geolocation system with applications to public health. In *AAAI HIAI Workshop*.

Duchi, J.; Hazan, E.; and Singer, Y. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR* 12:2121–2159.

Eisenstein, J.; Smith, N. A.; and Xing, E. P. 2011. Discovering sociolinguistic associations with structured sparsity. In *ACL*.

Gust, D. A.; Darling, N.; Kennedy, A.; and Schwartz, B. 2008. Parents with doubts about vaccines: which vaccines and reasons why. *Pediatrics* 122(4):718–725.

Hepburn, L.; Miller, M.; Azrael, D.; and Hemenway, D. 2007. The US gun stock: results from the 2004 national firearms survey. *Injury Prevention* 13(1):15–19.

Hu, Y.; Boyd-Graber, J.; Satinoff, B.; and Smith, A. 2014. Interactive topic modeling. *Machine learning* 95(3):423–469.

Jagarlamudi, J.; Daumé III, H.; and Udupa, R. 2012. Incorporating lexical priors into topic models. In *EACL*.

King, B. A.; Dube, S. R.; and Tynan, M. A. 2012. Current tobacco use among adults in the United States: Findings from the national adult tobacco survey. *American Journal of Public Health* 102(11):e93–e100.

Krosnick, J. A.; Judd, C. M.; and Wittenbrink, B. 2005. The measurement of attitudes. *The handbook of attitudes* 21–76.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 121–128.

Mimno, D., and McCallum, A. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *UAI*.

Myslín, M.; Zhu, S.-H.; Chapman, W.; and Conway, M. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15(8).

O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*.

Paul, M. J., and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*, 265–272.

Paul, M. J., and Dredze, M. 2013. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *HLT-NAACL*, 168–178.

Paul, M. J., and Dredze, M. 2015. SPRITE: Generalizing topic models with structured priors. *Transactions of the Association for Computational Linguistics (TACL)* 3:43–57.

Paul, M., and Girju, R. 2009. Cross-cultural analysis of blogs and forums with mixed-collection topic models. In *EMNLP*, 1408–1417.

Phan, X.; Nguyen, L.; and Horiguchi, S. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 91–100. New York, NY, USA: ACM.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. 2009. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.

Ramage, D.; Dumais, S. T.; and Liebling, D. J. 2010. Characterizing microblogs with topic models. In *ICWSM*.

Sheldon, D., and Dietterich, T. G. 2011. Collective graphical models. In *Neural Information Processing Systems (NIPS)*.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2951–2959.

Thacker, S. B., and Berkelman, R. L. 1988. Public health surveillance in the United States. *Epidemiologic reviews* 10:164–90.

Wallach, H.; Murray, I.; Salakhutdinov, R.; and Mimno, D. 2009. Evaluation methods for topic models. In *ICML*.

Yano, T.; Cohen, W.; and Smith, N. 2009. Predicting response to political blog posts with topic models. In *The 7th Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Zhu, X.; Blei, D.; and Lafferty, J. 2006. TagLDA: bringing document structure knowledge into topic models. Technical Report Technical Report TR-1553, University of Wisconsin.