Assessing Candidate Preference through Web Browsing History

Giovanni Comarela
Federal University of Viçosa

Ramakrishnan Durairajan
University of Oregon

Paul Barford
University of Wisconsin-Madison
and comScore, Inc.

Dino Christenson
Boston University

Mark Crovella
Boston University

ABSTRACT

Predicting election outcomes is of considerable interest to candidates, political scientists, and the public at large. We propose the use of Web browsing history as a new indicator of candidate preference among the electorate, one that has potential to overcome a number of the drawbacks of election polls. However, there are a number of challenges that must be overcome to effectively use Web browsing for assessing candidate preference—including the lack of suitable ground truth data and the heterogeneity of user populations in time and space. We address these challenges, and show that the resulting methods can shed considerable light on the dynamics of voters' candidate preferences in ways that are difficult to achieve using polls.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Web log analysis**; • **Computing methodologies** → *Supervised learning by classification*;

KEYWORDS

Candidate preference; browsing behavior; machine learning

ACM Reference Format:
Giovanni Comarela, Ramakrishnan Durairajan, Paul Barford, Dino Christenson, and Mark Crovella. 2018. Assessing Candidate Preference through Web Browsing History. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219884

1 INTRODUCTION

Understanding the candidate preference of voters leading up to a major election such as the 2016 U.S. presidential election is an important but difficult task. Polls are widely used, but have significant drawbacks; among other issues, a poll requires multiple days to complete, and hence cannot give insight into the short-term dynamics of vote choice, especially on a per-state level. Further, poll results are confounded by interviewer effects, question wording, and non-responsive or non-forthcoming subjects.

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219884

In light of the challenges presented by traditional polling, in this paper we undertake an exploration of an alternative approach for assessing candidate preference in the electorate. Our study examines the issues and potential benefits of approaching this problem by using passively collected records of user activity on the Web. In particular, we undertake the first study to look at the relationship between Web browsing behavior and election candidate preference. Our contributions are twofold: first, we elucidate the challenges involved in using browsing behavior to assess candidate preference and we present methods that can be used to overcome those challenges. Second, we show that using Web browsing behavior, it is possible to predict candidate preference with accuracy equivalent to state of the art polling, but with the additional advantage that prediction can be made on a fine-grain both spatially and temporally (i.e., at a state level on a day-to-day basis). State-level prediction is particularly important in the U.S. due to the winner-take-all nature of the electoral college system at the state level.

Our study is based on a large corpus of Web browsing behavior collected from participants who have opted-in to a major media measurement company (comScore, Inc.). These users are essentially analogous to Nielson families or participants in tracking polls. The dataset captures the browsing history of more than 100,000 people in the U.S. over a 56 day period just prior to the 2016 U.S. presidential election. Our high-level approach is to train a set of specially-designed classifiers at the start of the prediction period and then apply the classifiers on a day-by-day and state-by-state basis to assess fine-grained shifts in candidate preference over the remainder of the prediction period.

We show how to overcome the wide set of challenges inherent in applying this approach to browsing data. First, we address feature selection and show which aspects of Web browsing are most informative for prediction of candidate preference. Second, we address the challenge of learning a model for individual user behavior given that training labels on a per-user basis are not available. Rather, the best training data available consists of polling done at the per-state or per-county level. Third, we address the problem of making predictions on a state level, since training per-state models requires subdividing data, with a large subsequent loss of training accuracy. This problem is compounded by the fact that the amount of data per state varies tremendously due to both population density differences and geographical state sizes. Finally, we address the challenge of varying population composition as a function of day of the week. It is well known that user activity on the Web varies qualitatively based on day of the week. In order to make accurate predictions on a day to day basis, this effect must be corrected. For each of the above challenges, we clarify the nature of the problem and present new solutions.

Having overcome the above challenges, we then demonstrate the utility of the resulting method. We show that election results can be predicted on a per-state level with accuracy comparable to state of the art polling (with linear correlation of 0.94), and that the impact of an exogenous event (the release of the infamous 'Comey letter' on October 28, 2016) can be assessed for its impact on candidate preference on a fine-grained (day to day and state by state) basis.

In this paper we focus on the political preference; however we emphasize that our methods combined with Web browsing data and poll-like training data can be used to provide predictive capability in a wide variety of different scenarios. For instance, in place of candidate preference, one could assess opinions toward policy making, advertising impact, buying preferences, media consumption, leisure, etc. The advantage of considering browsing data is the possibility of analyzing important events and changes of opinion in (almost) real time, which is nearly impossible to do by relying on traditional techniques (e.g., polls).

2 RELATED WORK

2.1 Public Opinion Research Methods

Predicting elections requires insights into the citizenry's future behaviors, which in turn depend on understanding its political attitudes and preferences. Surveys, or polls, have become the most important and prevalent method of gauging public opinion and behaviors since their emergence in the 1930s. In a recent review of public opinion surveys, Berinsky [3] notes early observations of the power of polls in the minds of both the public and elites [10, 50], but also the inherent difficulty of the task [4, 15, 22, 30].

Indeed, in the last 75 years, researchers have come to understand the strengths and weaknesses of surveys, as well as the myriad of considerations necessary to properly use these instruments to understand political attitudes. Weisberg [51] summarizes the potential for error in surveys. Sources of errors include response accuracy (interviewer effects [e.g., 18, 53], question wording [see also 21], questionnaire issues [e.g., 11, 32, 41, 45, 49], and question nonresponse [e.g., 47]), respondent selection (unit nonresponse [e.g., 1, 14], sampling frames and error [e.g., 31, 36]), and survey administration (data editing, sensitive topics, and comparability effects).

Acknowledging the immense contribution of survey research to our understanding of political behavior and opinion, the vast potential for error suggests the possibility that other methods may shed additional insights. Indeed, other approaches are not uncommon in political science, including interpreting the partisan press, letters to newspapers, as well as public speeches and protests [see 22]. More recently political analysts have turned to the internet and social media, in particular, in search of correlates of public opinion and behavior. Related to our findings on the important role of social media sites, Fourney et al. [12] find that this is primarily where fake news stories were seen, and that aggregate voting patterns at both the state and county levels are strongly correlated with fake news site visits. Likewise, O'Connor et al. [38] note correlations between Twitter sentiment and public opinion polls during the 2008 presidential election, while Tumasjan et al. [48] find sentiments correspond to the parties' and politicians' political positions in a German federal election.

2.2 Election 2016 Forecasting

The difficulty of the election forecasting task in 2016 provides particular motivation for our methods. Election forecasting made waves in 2016 for the common perception that the polls *got it wrong*. Indeed, both the mean of the 10 election polls culled by *Real Clear Politics* on the day before the election as well as the "Polls-Plus" model forecasts on the *538* website suggested a Clinton victory with fairly high confidence. Notably, however, they were off in the popular vote only by a small margin, about .7% [5]. To be sure, other sites that rely heavily on daily updating of polls, such as the *Times's Upshot* and *Huffington Post*, as well as betting markets like *PredictIt*, generally held Clinton as the probable winner. By contrast, political science forecasting models primarily, but not exclusively, predict the popular vote.[1] The models of the popular vote based on political science fundamentals, like presidential approval, pre-campaign polls, incumbency, direction of the country, and the state of the economy did very well. Campbell et al. [5] notes that of the ten models seven missed Clinton's vote share by one percentage point or less and three missed it by less than half of a percentage point.

While what went wrong with the polls close to the election and poll aggregation models are still active areas of research, it should be clear that predicting electoral college outcomes is extremely challenging, not least because these are rare events with fluctuating voter turnout, and in close elections depend heavily on tight margins in a dozen or so *battleground states*.[2] This fact underscores the need to produce accurate predictions on a per-state basis, not just at the popular vote level. Given the limitations of polling as described here and evidenced in the 2016 election, it is clear that additional information sources and methods may be helpful in addressing this challenge.

2.3 Machine Learning in the Study of Politics

Grimmer [19] notes that social scientists have primarily used machine-learning for characterizing and interpreting data, as opposed to using it for prediction. In particular, social scientists have used machine learning techniques to understand the latent qualities of text [e.g., 6, 20, 26, 43]. However, supervised methods are gaining in popularity. The discipline's focus on estimating complex causal effects has led to applications here [e.g., 23, 24]. Moreover, these techniques are making inroads to political predictions as well. For example, Kennedy et al. [29] develop prediction models for more than 500 elections across 86 countries based on polling data. Of particular relevance to our work, Kaufman et al. [27] use AdaBoosted decision trees to predict Supreme Court decisions, as well as the Democratic Party's county-level vote share from census data on age, income, education, and gender. Montgomery and Olivella [37] apply related methods to U.S. elections to predict campaigns' decisions to go negative, and to estimate the attitudes and behaviors of demographic subgroups. Ultimately, they foresee decision tree models as being most useful for analyzing complex data generating processes where the goal is prediction (rather than theory testing).

[1] For a more thorough engagement with forecasting in 2016 see the collection of articles in the *PS: Political Science & Politics* "Symposium: Forecasting the 2016 American National Elections" https://www.cambridge.org/core/journals/ps-political-science-and-politics/issue/4A42B1C2814C474155929789CCBAD7D5.

[2] For a review of the history of election forecasting see Lewis-Beck [35].

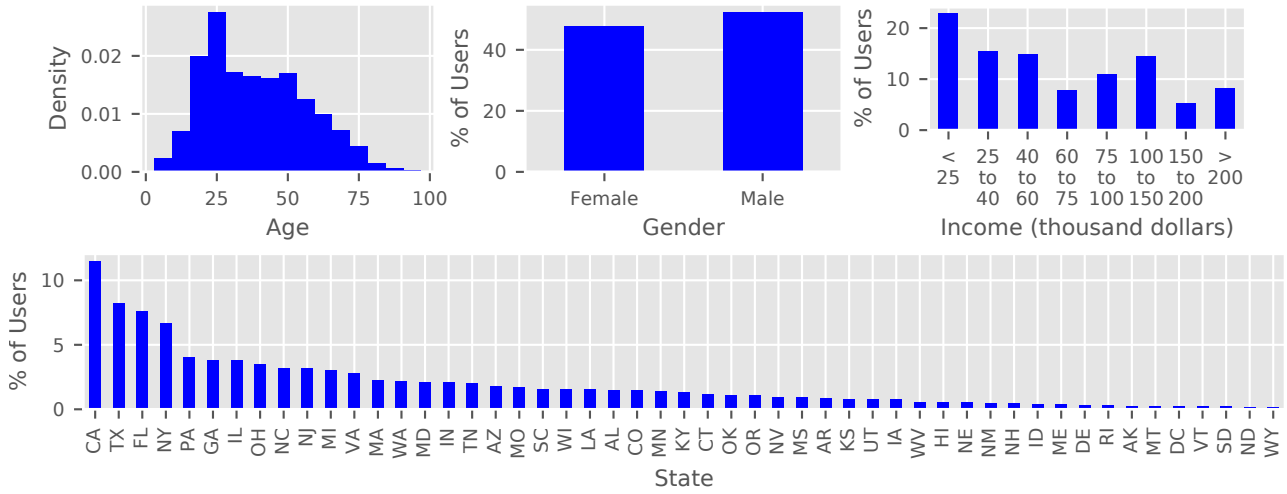

Figure 1: Demographics of User Population Studied

2.4 Analyzing Social Media and Networks

Surveys during the 2016 campaign showed that 27% of respondents found either social media or news websites to be the most useful way to get election information [17]. Accordingly, there has been substantial work using data from social media (Twitter and Facebook) to assess and predict political opinions, sentiments, and preferences [2, 7, 9, 16, 34, 40, 46, 48]. When supervised learning is employed, ground truth labels are obtained from small subsets of users who self-declare (e.g., [40]), or from small sets of manually-labelled users (e.g., [9, 16]).[3]

These studies contrast with ours in a number of ways. As we describe in Section 4.2, our method does not require per-user ground-truth labels, and so can make use of much larger training sets. Second, due to the nature of social media, previous methods generally focus on textual, linguistic, sentiment, and network analysis; a number of studies have outlined the weaknesses of these methods for predicting electoral outcomes [13, 25]. In contrast, we show how to use knowledge of Web sites visited to make predictions, which opens up a potentially simpler and larger body of data for analysis.

In summary, our methodological approach has a number of strengths. Foremost, perhaps, it relies on actual behavior, sidestepping the potential response accuracy problems that come with asking respondents to interpret questions or recall their own behaviors. Moreover, contrary to the complexity of inferring sentiment from text on social media, blogs, or the likes, our approach looks at sites visited, a straightforward and simple behavior, but one we find to be extremely informative. We overcome the challenge of obtaining per-user ground truth by developing a learning approach that can use polling data for training. Finally, while our sample may not be perfectly generalizable to the electorate, it is *big*, thereby providing substantial demographic and geographic heterogeneity, allowing predictions to be made on a fine grain both spatially and temporally.

[3]We note that there is some evidence that reported accuracies of such methods are overstated [7].

3 DATA

Our study is based on Web browsing data provided by comScore's global desktop user panel.[4] The comScore panel is comprised of of over 2 million residential users who are compensated for their participation. When a user registers to participate in the panel, they voluntarily provide their home address along with basic demographic information such as age, sex and income. Upon completion of registration, they download and install monitoring software that is active whenever they browse the Web. The software captures a variety of data about their browsing activity including the URLs transmitted in Web requests. We note that panelist privacy and security are paramount concerns for comScore; details are at [8].

The data that we use in this study consists of all Web browsing activity of US comScore panelists over a 56-day period from September 9, 2016 to November 3, 2016. Table 1 provides an overview of the corpus, after the data reduction described below.

The demographics of our user population are shown in Figure 1. We note that it is not necessary for the demographics of the user population to precisely match the demographics of the underlying populations, since we are training per-state models that compensate for variations in user population composition across states (as discussed in Section 4.3). However, the figure shows the extreme imbalance in population sizes across states, which presents one of the key challenges to our approach (and which we show how to address in Section 4).

Data reduction. The total size of the data corpus was approximately 1.6 TB in compressed form. Data reduction and feature extraction were performed using Apache Hadoop on a 19-node cluster.

While the panel data includes full URLs for data requested by panelists, we discard the portion of the URL to the right of the site's domain name. Our initial explorations indicated that more detailed portions of URLs did not provide enough signal in general to improve our estimates. However, as described in Section 4.1, we

[4]comScore is a global media measurement company – www.comscore.com.

Total unique panelists	~120 k
Total unique URLs	~70 M
Total unique sites	~380 k
Avg. URLs (non unique) visited per day per user	~140

Table 1: Overview of the Web browsing data that is the basis for our study. The data was provided by US participants in comScore's global desktop user panel and was collected from September 9, 2016 to November 3, 2016.

found other features such as HTTP referral headers to be important. We also removed ads, by filtering out URLs matching those in easylist.to, and mapped the time (originally UTC) of each request to the local time of the correspondent panelist.

4 METHODS

In this section we describe the four main challenges to be overcome in applying our approach, and the new approaches we developed to address them.

4.1 Feature Selection

Our first challenge is to extract a set of features that effectively captures user preference from the user browsing logs described in Section 3. We considered two sets of features: Web domains visited, and HTTP referral information. For each user in the dataset, on each given day, we created four feature vectors, each of which was normalized to unit sum:

- **Alexa.** The frequency with which the user visited each one of the top 500 sites in the US, according to `Alexa`.[5] We considered this our baseline feature vector.
- **Social Media.** The frequency with which each user visited sites when **referred by** social media sites (`facebook.com` and `twitter.com`). Only visits to the top 100 most visited sites (from social media sites) according to our dataset were considered.
- **Search Engine.** As for **Social Media**, but considering referrals from large search engines (`google.com` and `bing.com`). Only visits to the top 100 sites referred from search engines were considered.
- **None.** Similar to the last two, but considering only sites without any referral information. Also for top 100 sites.

We chose the top 500 sites from `Alexa` because it was the maximum available at a country level. We considered only 100 domains for the referral vectors to avoid adding many sparse features. Then, we conducted experiments in order to decide which of these vectors were more appropriate to our problem. We combined the base feature vector, *Alexa*, with the other 3 individually. In initial experiments (similar to the ones in Section 5.2), we observed significant improvements when combining the *Alexa* with the *Social Media* vector. We also observed that no significant improvements were reached when augmenting the *Alexa* vector with *Search Engine* or *None*.

We conclude that, in general, people's Web browsing behavior when referred from social media sites is a better indication of their candidate preference than sites visited from search engines or sites visited by directly typing URLs on the browser. Hence, in the remainder of this work, feature vectors representing users have size 600, with the first 500 components holding the distribution of visits to the top 500 sites in the US, and the last 100 components reflecting the distribution of visits to sites referred by social media. These feature vectors are computed each day for each user having activity on that day.

[5] https://aws.amazon.com/alexa-top-sites/

Algorithm 1: `EM-training`

Data: $\mathbf{U}$, R, B, $\mathcal{A}$

$L' \leftarrow$ `InitLabels`(U, R, B)

repeat

  $L'' \leftarrow L'$

  $\Theta \leftarrow$ train $\mathcal{A}$ on $(\mathbf{U}, L')$

  foreach *region* r **do**

    $t_r \leftarrow$ percentile $(1 - B(r))$ of $P(L(u) = 1|\Theta)$, $\forall u \in r$

  foreach *row* u **do**

$$L'(u) = \begin{cases} 1, & \text{If } P(L(u) = 1|\Theta) \geq t_{R(u)} \\ 0, & \text{otherwise} \end{cases}$$

until $L' \approx L''$;

return Θ

4.2 Learning without Individual Labels

The most significant challenge to overcome to use supervised learning for political preference is that at the individual level, labels (i.e., candidate preference) are generally not known. As described in Section 2, previous work has been limited to very small training set sizes due to the need to obtained self- or manually-labeled user preferences.

In contrast, we start from the observation that what is in general available is polling data – the fraction of voters from a given region (e.g, city, county, state, and country) that prefer a given candidate *at a given point in time*. Such data can be interpreted as providing knowledge that *some* number of users has a particular label, without identifying *which* users have that label.

We therefore consider the problem of starting from known polling measurements at a specific time, and predicting individual user preference *forward* from that point (with fine resolution in time and space). To address this we design an algorithm motivated by the general *expectation maximization* (EM) principle. At a high level, we adopt a data augmentation strategy of assigning a putative label to each user. In the E-step, we assign each user a label, giving out labels in proportion to observed polling data. In the M-step, we estimate parameters of our classifier, resulting in a new ranking of users with respect to label probability. The resulting algorithm jointly learns both putative user labels and also a user classifier, given regional statistics and user features.

To make our approach concrete, we first present basic definitions and our EM-training algorithm in a general setting. Then, we show how we instantiate our algorithm to solve our problem. At the end of this section, we also describe a simpler approach that we use as a baseline for comparison.

Our EM-training algorithm uses the following definitions:

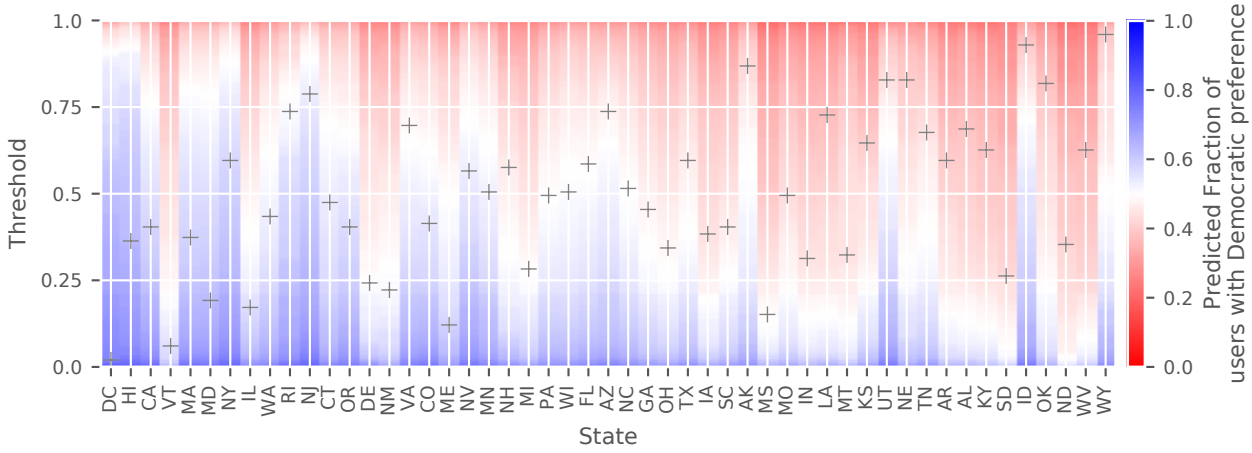

Figure 2: No single threshold can correctly predict the fraction of users preferring Democrats in each state.

Algorithm 2: `InitLabels`

Data: U, R, B

1 **foreach** *row* u **do**

2 $\quad L'(u) = \begin{cases} 1, & \text{If } B(R(u)) \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$

3 **return** L'

- **U**: data matrix having users as rows and features as columns;
- $R(u)$: region to which user u belongs to;
- $L(u)$: true label of user u (unknown);
- $L'(u)$: inferred label of user u, according to model parameterized by Θ;
- $B(r)$: known fraction of users in region r having label 1;
- $P(L(u) = 1|\Theta)$: estimated probability that user u has label 1, according to model parameterized by Θ;
- $\mathcal{A}$: learning algorithm yielding Θ that predicts $P(L(u) = 1|\Theta)$.

Our goal then is, given **U**, R, B and $\mathcal{A}$, to obtain a model Θ that maximizes

$$\sum_u \mathbb{1}_{[L'(u)=L(u)]} - \sum_u \mathbb{1}_{[L'(u)\neq L(u)]} \quad (1)$$

where $\mathbb{1}_{[.]}$ is the indicator function. In other words, we aim at obtaining a model that can accurately predict L, even though it is unknown.

Our method is shown in Algorithm 1. It starts by obtaining an initial label assignment L' in line 1. The goal this step is to have labels that allow $\mathcal{A}$ to start the learning process in the right direction; we discuss the function `InitLabels` later. Next, there is an iterative process that: *i)* uses the current label assignment to train a new, and improved, model; *ii)* uses the new model in order to obtain a better label assignment. The process is repeated until the current and new label assignments are approximately the same (e.g., by changing less than 1%).

The M-step is described in line 4 and it consists basically of training $\mathcal{A}$ on matrix **U** and labels L'. The two loops between lines 5 and 8 are the E-step, responsible for creating the new label assignment. The idea is to take advantage of the fact that we know that each region r has a fraction $B(r)$ of users with label 1. To that end, we assign label 1 to the users with highest probability of having label 1. That is, we assign label 1 to all users in region r that are above the $1 - B(r)$ percentile of the $P(L(u) = 1|\Theta)$ values. During the execution of the main loop of the algorithm, in each region r only a fraction $B(r)$ of users has label 1. Our hypothesis is that upon convergence, these users are in fact the mostly likely ones to have such label.

A key part of Algorithm 1 is the label initialization (line 1). The challenge in initialization is that there is no information about which users within a region are more or less likely to have label 1. Therefore, if we want to distribute the such label to a fraction $B(r)$ of the users in region r, we cannot do it better than in a random fashion. To overcome this problem, our strategy, presented on Algorithm 2, uses the idea of assigning label 1 to all users of a region for which $B(r) \geq 0.5$ and label 0 to all users of regions with $B(r) < 0.5$. Two advantages arise from considering this simple initialization. First, it eliminates adding any randomness to the algorithm, avoiding different output models for different runs with the same input. Second, the value of Equation 1, our desired objective function, is higher with this initialization than it is using the random approach. In order to observe such fact, consider a region r with n users. Then, it is possible to show that the expected value of Equation 1 under the random approach is $n(2B(r)-1)^2$. Also, it can be shown that the same value, when using Algorithm 2, is $n(2B(r) - 1)$ if $B(r) \geq 0.5$ and $n(1-2B(r))$ when $B(r) < 0.5$, which is larger than $n(2B(r)-1)^2$ in the $(0, 1)$ interval.

To instantiate this algorithm, in this paper we created the matrix **U** with rows representing users in the comScore dataset and features as described in Section 4.1; for the set of regions we used the US states (plus DC), totalling 51 regions; for the function B, we used state-level polls from the 538 site as reported on September 10, 2016;[6] and as learning algorithm $\mathcal{A}$ we used regularized logistic regression. Finally, we also fix label 1 to mean a preference for

[6] https://projects.fivethirtyeight.com/2016-election-forecast/

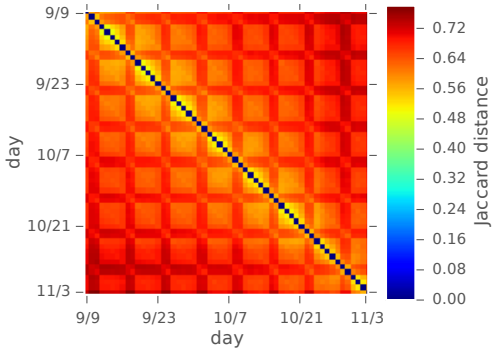

Figure 3: Changes in the set of users. Color represents the Jaccard distance between sets of users for all possible pairs of days.

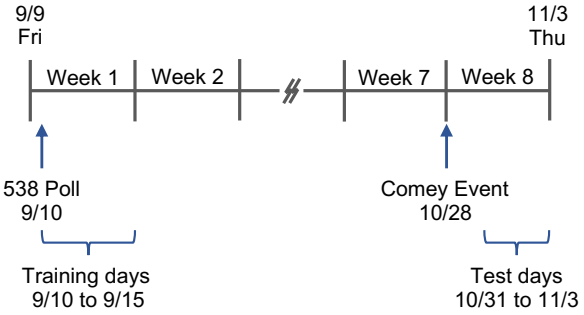

Figure 4: Dataset timeline. Highlighting: data retained for training the model and for predictions; day of poll data; and date of exogenous event (i.e., Comey letter).

Hilary Clinton (or for Democratic) and 0 to mean a preference for Donald Trump (or for Republican).

We note that the problem of learning with label proportions has been addressed in the literature (e.g. [33, 39, 42, 52]). In general, we find these methods are slow and do not scale easily to the size of data needed to assess opinions on a fine-grain. For the same reason (i.e., scalability) we were not able to replace logistic regression with more advanced classifiers (e.g. support vector machines) in Algorithm 1. Furthermore, as we show in the next section, logistic regression allows for natural incoporation of a single state-specific parameter in ways that more sophisticated methods do not. Ultimately though, we emphasize that the main focus of this work was not on solving the machine learning problem itself, but on exploring the potential for Web browsing behavior to shed light on candidate preference. For this problem, as it will be shown in the next sections, our EM-based approach using logistic regression is quite effective.

4.3 Population Heterogeneity: Spatial

Our third methodological challenge concerns the spatial heterogeneity of user populations. Diving up users into state-level populations, we find that there are significant differences in the relationship between user (browsing) behavior and candidate preference across different populations.

One straightforward approach to address this problem is simply to train individual state-level models using Algorithm 1. However this means that each model has much less data to use for training, leading to greater variance. This problem is made much worse by the the extreme imbalance between state-level populations. As shown in Figure 1, there are order of magnitude differences between the sizes of state level populations in our data. As a result, there is simply not enough data in many states to estimate the 500 parameters of our logistic regression model.

However, we find that there *is* enough state-level data in each case to estimate *one* additional model parameter. Combined with the fact that using logistic regression as a classifier requires setting a classification threshold, this gives us a natural way to combine the strength of our nationwide population for training *most* model parameters, with state-level polls for setting *one* state-level classification threshold.

The need for state-level models can be seen in Figure 2. This figure shows how setting the classification threshold affects the predictions of our logistic regression model in each state. Colors correspond to the fraction of users that were predicted to have Democratic preference, as a function of the given classification threshold (using data from September 10, 2016). The plus symbol in each column is the learned threshold for best predicting the 538 polling data in that state on September 10.

The figure shows how variable the relationship between browsing behavior and preference is. The correct threshold varies considerably across states. This means that browsing behavior that our model would consider to be to a given degree 'Democratic' in one state could be considered much less (or more) 'Democratic' in another state. Our hybrid approach of having 51 models, each using 500 shared, nationwide parameters and one state-level parameter, addresses this problem while allowing us to make maximal use of data for training.

4.4 Population Heterogeneity: Temporal

The final challenge we need to overcome concerns population variation over time. Not all users are active on each day. To assess the effects of population changes over time, we proceed as follows. First, for each day, we compute the set of users active in that day. Next, for all pairs of days, we compute the Jaccard distance between the corresponding sets of users.

The results are shown in Figure 3. Two effects are present. First, we note changes on a long time-scale (weeks or more) that happen due to new users joining or leaving the comScore panel. This effect is relatively weak and we find that it does not affect prediction accuracy to a significant degree.

The more significant effect is a weekly pattern, where the sets of users on weekdays are similar, sets of users on weekends are similar, and there is a large difference between weekday and weekend users. This trend is probably due to the different habits that people have during week days (mostly working days) and weekends. Interestingly, we find that models trained on weekday users tend to predict that weekend users have greater preference for the Democratic candidate than they really do. Hence we adopt a simple clustering strategy, in which we partition all our data into two (weekday and weekend) sets, and train and predict on the two sets separately.

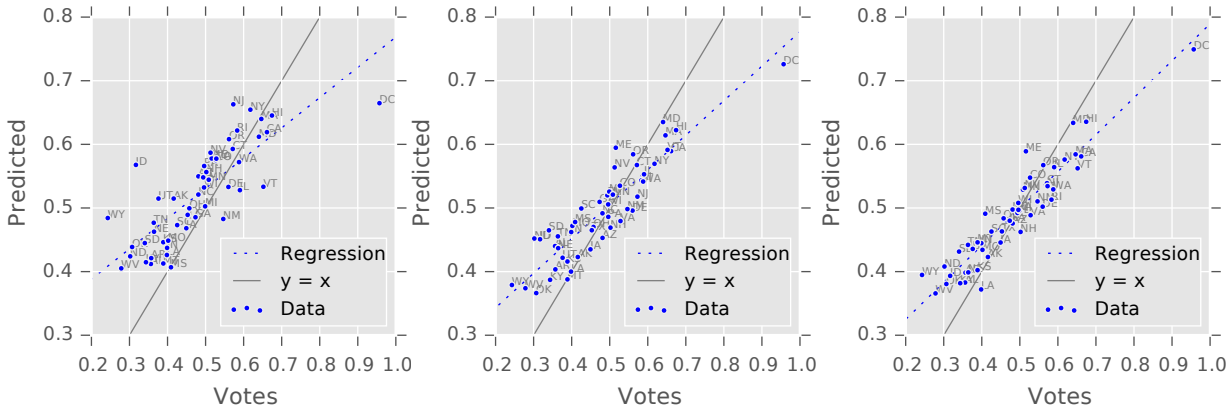

Figure 5: Fraction of support for Democrats. *Left:* baseline model ($\rho = 0.80$); *Center:* baseline plus threshold refinement ($\rho = 0.91$); *Right*: baseline plus EM plus threshold refinement ($\rho = 0.94$).

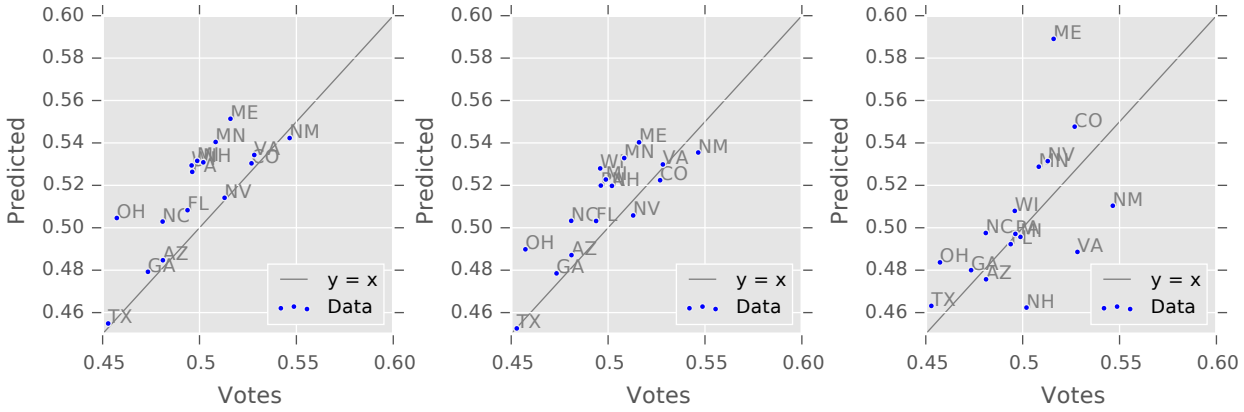

Figure 6: Fraction of support for Democrats. *Left:* Election results vs. 538 on September 10; *Center:* Election results vs. 538 on November 7; *Right*: Election results vs. our final predictions.

5 RESULTS

In this section we show the performance of our method in practice. We start by presenting how we split our data in order to avoid overfitting (in Section 5.1). In Section 5.2, we discuss how the learning algorithms presented in Section 4 and how our state-level predictions compare to polls and actual election results. Finally, in Section 5.3 we show the impact of exogenous events on user's browsing behavior, by focusing on a case study, i.e., the Comey letter announcement on October 28.

5.1 Putting it All Together

Figure 4 shows how we divided the dataset for purposes of training and testing. We retained data from the first week for training. We used data from September 10 and 11 for the weekend model, and from September 12 to 15 for the week model. For testing, i.e., correlating with election results, we used the four last days of the dataset (i.e., from October 31 to November 3). Also for the purpose of training the model, we used 538 polls from September 10, both for estimating $B(r)$ (fraction of Hilary supporters in each state) and tuning the per-state logistic regression thresholds.

In the Figure, we also point to October 28, i.e., the date of the Comey letter announcement, which is analyzed on Section 5.3. In this context, we used the model trained with data from the first week to predict the support to the candidates in the remaining 7 weeks. Weekend and week days predictions were performed using the correspondent models. Moreover, we define the time after (inclusive) October 28 as *after* Comey, and from September 16 to October 27 as *before* Comey.

5.2 Prediction Accuracy

In this section we demonstrate the methods presented in Section 4 and assess the results. Since users' true labels are not known, we cannot analyze the models in terms of accuracy, precision, or recall. Hence, we use instead the correlation between our state-level predictions and the actual election results.[7]

Figure 5 shows the contribution of the various aspects of our method. The left hand plot shows our 'baseline' model, which only

[7]Note that in all the results we report, we exclude from consideration all polling and election results for candidates other than Clinton and Trump, meaning that percentages for the two principal candidates always sum to 100%.

trains a single logistic regression model and uses statewide majority to label users. The center plot shows the improvement due to using 51 state-level models; the baseline model is adjusted to set thresholds on a per-state basis, leading to an improvement of 0.11 in linear correation. Finally, the right hand plot shows the improvement due to incorporating our EM-training algorithm, leading to an additional improvement of 0.03 in linear correlation. These results confirm the importance of state-level models and the utility of our EM-training approach.

To study predictions in detail, we look at the battleground states, which we define here as states in which the final vote tally was between 45% and 55% for Clinton. Figure 6 compares our models' predictions against the polls as reported by the 538 site. On the left of the figure we show comparison of the vote with the polls as of September 10; in the center we show comparison to the polls as of November 7; and on the right we show comparison with our models' predictions.

Overall the browsing based models show more absolute error compared to the reference values (the final vote on November 11) than do the polls. However, the figures show that the polls tended to generally overpredict the fraction of votes for Clinton, while the browser-based models make predictions that are more equally balanced between over-prediction and under-prediction for Clinton. In fact the mean prediction error over the battleground states for our models is 0.005, while for the polls it is 0.018 (September 10) and 0.012 (November 7). This translates to a more accurate prediction of the state-level outcomes (and hence the national election). Specifically on September 10, polls predicted incorrect outcomes on 7 states (MI, PA, WI, FL, NC, OH, and IA); on November 7, polls were wrong on 5 states (MI, PA, WI, FL, and NC); while our models' estimates are wrong on only 3 states (VA, NH, and WI).

These results suggest that using Web browsing behavior for prediction of candidate preference, it is difficult to do better than polls in absolute terms. However the accuracy of these models is at least comparable, and since they use an entirely different source of (passively collected) information for estimation they may be able to avoid some of the bias that creeps into polling results.

5.3 Assessing an Exogenous Event

A significant promise of our method is the potential for fine-grained analysis of user preference, potentially leading to insights about how various factors influence the electorate. To illustrate this promise, we consider the widely-discussed release of the so-called 'Comey letter' on October 28, 2016: in early July 2016, FBI Director James Comey stated that the FBI would not be bringing charges against Democratic candidate Hillary Clinton regarding an issue of her handling of certain classified information while she was Secretary of State. Then, on October 28, Director Comey announced that the FBI was opening a review of new evidence in the case. This caused immense controversy in the press, and many commentators (including President Bill Clinton) attributed Donald Trump's subsequent victory to this 'October surprise' [44].

To assess the impact of the 'Comey letter' we use our models to make daily predictions of the preferences of the electorate in the two weeks preceding Nov 4. As a comparison, we take the daily average of predicted electorate preference over the six-week period

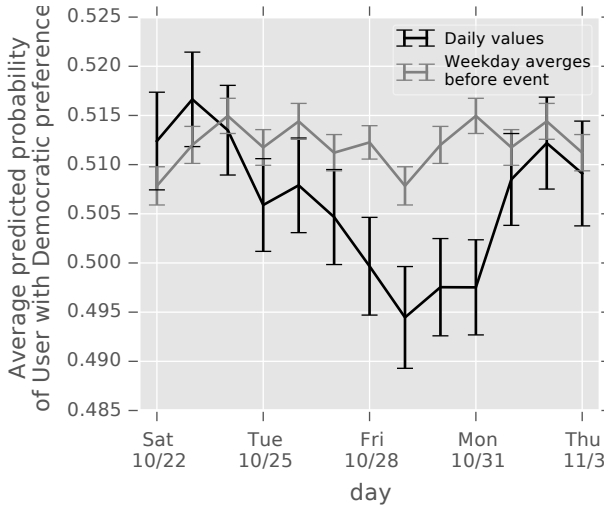

Figure 7: Electorate preference surrounding the 'Comey letter' (October 28) compared to previous weekly average.

before October 28 (see Figure 4). The figure shows that during the period October 28 to October 31, there was a dramatic and statistically significant drop in Democratic preference across the electorate. This shows how the temporal impact of such an event can be assessed on a day to day basis.

Interestingly though, the figure also shows that Democratic preference was headed downward *notably in advance* of the release of the 'Comey letter.' A significant drop in Democratic preference appears starting around October 25. This suggests that the Comey letter may have accelerated an effect that was already in progress.

Indeed, this analysis was puzzling in light of the narrative in the popular press. However, as our paper was in final preparation, an extensive and detailed study of the 2016 election polls was released [28]. Quoting from that study (p. 22):

> The evidence for a meaningful effect on the election from the FBI letter is mixed at best. Based on Figure 6, it appears that Clinton's support started to drop on October 24th or 25th. October 28th falls at roughly the midpoint (not the start) of the slide in Clinton's support.

We conclude that the fine-timescale analysis afforded by our method adds nuance to the popular understanding of the 'Comey letter' effect, while being strikingly in agreement with the most recent detailed analysis based on polling data.

Augmenting this fine-timescale analysis, we can also look at the 'Comey letter' effect on a fine spatial scale. State-level analysis is important because of the number of states, many midwestern, in which predictions for a Clinton win based on polling turned out to be incorrect.

Figure 8 shows the difference in preference for the Democratic candidate before versus after the 'Comey letter.' Specifically, it shows the average Democratic preference on Fri-Mon across the six weeks prior to October 28, minus the average Democratic preference on Fri-Mon subsequent to October 28. Due to the smaller sample sizes involved at the state level, individual state level differences

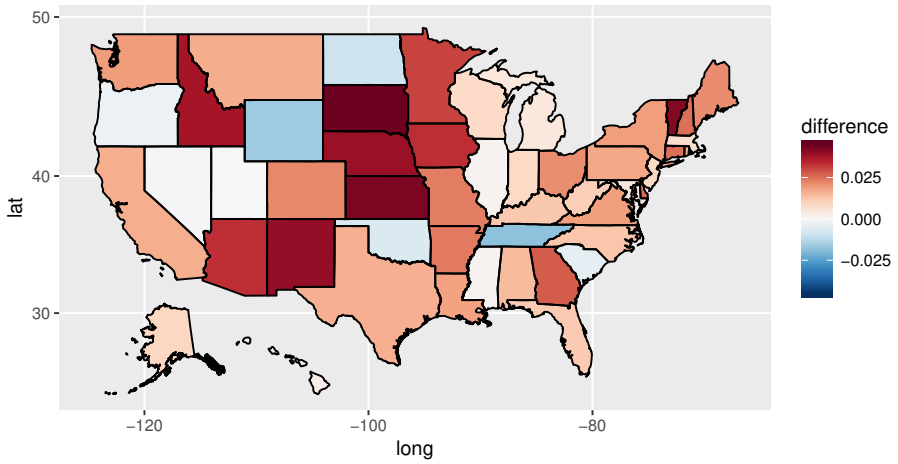

Figure 8: Impact of the 'Comey letter' at the state level.

do not rise to the level of statistical significance, but nationwide trends are nonetheless clear.

The figure shows that the popular shift away from Clinton was stronger in the west and midwest of the country. In some cases, outliers can be understood as a result of small sample sizes (Wyoming, Oklahoma, North Dakota). However, overall it is clear that there were moves away from Clinton in some states with extremely close margins - NH (.4%), MN (1.5%), AZ (3.9%), PA (1.2%), and thus the 'Comey letter' may have been a determinative factor in the election as a whole.

6 CONCLUSION

In this paper we have presented the first method that uses the history of user visits to Web sites to assess individual preference for political candidates. In doing so we make a number of contributions. First, we pinpoint the challenges to be overcome in realizing this goal, chief among them dealing with temporal and regional heterogeneity in user populations, as well as overcoming the lack of individual-level ground truth labels for training. With respect to the latter, we develop a new method allowing us to train a user-level classifier using only aggregate (statewide polling) data. Second, we observed that "social referrals", i.e., visits to sites originated from social media, are more important to infer candidate preference than those originated from other sources, such as search engines and URLs directly typed into the browser. Third, we show the power of using Web browsing behavior for assessing candidate preference, particularly in terms of day-to-day and state-to-state level predictions that elucidate the impact of exogenous effects such as the 'Comey letter.'

Our results suggest that access to browsing data gives considerable power to assess the preferences of the electorate. With respect to understanding candidate preference, we believe that further use of Web browsing data is likely to uncover additional insights about the impact of campaign strategies, candidate speeches and visits, and political ad campaigns. More broadly, we believe that the methods we present here are not fundamentally limited to studying candidate preference and can be applied in a wide range of settings in which we have access to consistent polling to train with, including issue preference, attitudes, and identification.

ACKNOWLEDGEMENTS

This material is based upon work supported by NSF grants IIS-1421759, CNS-1618207, and CNS-1703592 and by AFRL grant FA8750-12-2-0328. The authors thank comScore for making their panel data available for this study.

REFERENCES

[1] J Scott Armstrong and Terry S Overton. 1977. Estimating nonresponse bias in mail surveys. *Journal of marketing research* (1977), 396–402.
[2] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[3] Adam J Berinsky. 2017. Measuring Public Opinion with Surveys. *Annual Review of Political Science* 20 (2017), 309–329.
[4] Herbert Blumer. 1954. What is wrong with social theory? *American sociological review* 19, 1 (1954), 3–10.
[5] James E Campbell, Helmut Norpoth, Alan I Abramowitz, Michael S Lewis-Beck, Charles Tien, Robert S Erikson, Christopher Wlezien, Brad Lockerbie, Thomas M Holbrook, Bruno Jerôme, et al. 2017. A recap of the 2016 election forecasts. *PS: Political Science & Politics* 50, 2 (2017), 331–338.
[6] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*. 288–296.
[7] Raviv Cohen and Derek Ruths. 2013. Classifying political orientation on Twitter: It's not easy!. In *ICWSM*.
[8] comScore, Inc. 2018. Panelist Privacy Statement. http://www.comscore.com/About-comScore/Privacy.
[9] 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) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. IEEE, 192–199.
[10] Philip E Converse. 1987. Changing conceptions of public opinion in the political process. *The Public Opinion Quarterly* 51 (1987), S12–S24.
[11] Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57, 1 (2007), 163–173.
[12] Adam Fourney, Miklos Z Racz, Gireeja Ranade, Markus Mobius, and Eric Horvitz. 2017. Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 2071–2074.
[13] Daniel Gayo Avello, Panagiotis T Metaxas, and Eni Mustafaraj. 2011. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

[14] Andrew Gelman, Sharad Goel, Douglas Rivers, David Rothschild, et al. 2016. The mythical swing voter. *Quarterly Journal of Political Science* 11, 1 (2016), 103–130.

[15] Benjamin Ginsberg. 1986. *The captive public: How mass opinion promotes state power.* New York: Basic Books.

[16] Jennifer Golbeck and Derek Hansen. 2014. A method for computing political preference among Twitter followers. *Social Networks* 36 (2014), 177–184.

[17] Jeffrey Gottfried, Michael Barthel, Elisa Shearer, and Amy Mitchell. 2016. The 2016 presidential campaign–A news event that's hard to miss. *Pew Research Center* 4 (2016).

[18] Pamela Grimm. 2010. Social desirability bias. *Wiley international encyclopedia of marketing* (2010).

[19] Justin Grimmer. 2015. We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics* 48, 1 (2015), 80–83.

[20] Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis* 21, 3 (2013), 267–297.

[21] Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology.* Vol. 561. John Wiley & Sons.

[22] Susan Herbst. 1993. *Numbered voices: How opinion polling has shaped American politics.* University of Chicago Press.

[23] Daniel W Hill and Zachary M Jones. 2014. An empirical evaluation of explanations for state repression. *American Political Science Review* 108, 3 (2014), 661–687.

[24] Kosuke Imai and Aaron Strauss. 2010. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis* 19, 1 (2010), 1–19.

[25] Andreas Jungherr, Pascal Jürgens, and Harald Schoen. 2012. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment". *Social science computer review* 30, 2 (2012), 229–234.

[26] Aaron Kaufman. 2018. Estimating the Partisan Bias of Survey Questions. (2018). Working paper.

[27] Aaron Kaufman, Peter Kraft, and Maya Sen. 2018. Improving Supreme Court Forecasting Using Boosted Decision Trees. (2018).

[28] Courtney Kennedy, Mark Blumenthal, Scott Clement, Joshua D. Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers, Lydia Saad, and G. Evans Wittand Christopher Wlezien. 2018. An Evaluation of the 2016 Election Polls in the United States. *Public Opinion Quarterly* (February 3 2018).

[29] Ryan Kennedy, Stefan Wojcik, and David Lazer. 2017. Improving election prediction internationally. *Science* 355, 6324 (2017), 515–520.

[30] Valdimar Orlando Key. 1961. Public opinion and American democracy. (1961).

[31] L Kish. 1965. *Survey Sampling.* Wiley, New York.

[32] Jon A Krosnick, Neil Malhotra, and Urja Mittal. 2014. Public misunderstanding of political facts: How question wording affected estimates of partisan differences in birtherism. *Public opinion quarterly* 78, 1 (2014), 147–165.

[33] Hendrik Kück and Nando de Freitas. 2012. Learning about individuals from group statistics. *CoRR* abs/1207.1393 (2012).

[34] Huyen T Le, GR Boynton, Yelena Mejova, Zubair Shafiq, and Padmini Srinivasan. 2017. Revisiting The American Voter on Twitter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, 4507–4519.

[35] Michael S Lewis-Beck. 2005. Election forecasting: principles and practice. *The British Journal of Politics & International Relations* 7, 2 (2005), 145–164.

[36] S Lohr. 1999. *Sampling: Design and Analysis.* Number v. 1 in Sampling: Design and Analysis. Duxbury Press.

[37] Jacob M Montgomery and Santiago Olivella. 2016. Tree-based models for political science data. *American Journal of Political Science* (2016).

[38] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, Noah A Smith, et al. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *Icwsm* 11, 122-129 (2010), 1–2.

[39] Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. 2014. (Almost) No Label No Cry. In *International Conference on Neural Information Processing Systems.*

[40] Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. *Icwsm* 11, 1 (2011), 281–288.

[41] Philip M Podsakoff, Scott B MacKenzie, Jeong-Yeon Lee, and Nathan P Podsakoff. 2003. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of applied psychology* 88, 5 (2003), 879.

[42] Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, and Quoc V. Le. 2008. Estimating Labels from Label Proportions. In *International Conference on Machine Learning.* 776–783.

[43] Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54, 1 (2010), 209–228.

[44] Nate Silver. 2016. The Comey letter probably cost Clinton the election. https://fivethirtyeight.com/features/the-comey-letter-probably-cost-clinton-the-election/.

[45] Paul M Sniderman and Sean M Theriault. 2004. The structure of political argument and the logic of issue framing. *Studies in public opinion: Attitudes, nonattitudes, measurement error, and change* (2004), 133–65.

[46] Stefan Stieglitz and Linh Dang-Xuan. 2013. Social media and political communication: a social media analytics framework. *Social Network Analysis and Mining* 3, 4 (2013), 1277–1291.

[47] Roger Tourangeau and Ting Yan. 2007. Sensitive questions in surveys. *Psychological bulletin* 133, 5 (2007), 859.

[48] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment.. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.*

[49] Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science* 211, 4481 (1981), 453–458.

[50] Sidney Verba. 1996. The citizen as respondent: sample surveys and American democracy presidential address, American Political Science Association, 1995. *American Political Science Review* 90, 1 (1996), 1–7.

[51] Herbert F Weisberg. 2009. *The total survey error approach: A guide to the new science of survey research.* University of Chicago Press.

[52] Felix X. Yu, Dong Liu, Sanjiv Kumar, Tony Jebara, and Shih-Fu Chang. 2013. ∝SVM for Learning with Label Proportions. In *International Conference on Machine Learning.*

[53] Daniel John Zizzo. 2010. Experimenter demand effects in economic experiments. *Experimental Economics* 13, 1 (2010), 75–98.