Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

"Voters of the Year": 19 Voters Who Were Unintentional Election Poll Sensors on Twitter

William Hobbs, Lisa Friedland, Kenneth Joseph, Oren Tsur, Stefan Wojcik, David Lazer
Network Science Institute, Northeastern University
Boston, Massachusetts 02115
w.hobbs@northeastern.edu, lisadfriedland@gmail.com, josephkena@gmail.com,
o.tsur@neu.edu, stefan.wojcik@colorado.edu, d.lazer@neu.edu

Abstract

Public opinion and election prediction models based on social media typically aggregate, weight, and average signals from a massive number of users. Here, we analyze political attention and poll movements to identify a small number of social "sensors" – individuals whose levels of social media discussion of the major parties' candidates characterized the candidates' ups and downs over the 2016 U.S. presidential election campaign. Starting with a sample of approximately 22,000 accounts on Twitter that we linked to voter registration records, we used penalized regressions to identify a set of 19 accounts (sensors) that were predictive for the candidates poll numbers (5 for Hillary Clinton, 13 for Donald Trump, and 1 for both). The predictions based on the activity of these handfuls of sensors accurately tracked later movements in poll margins. Despite the regressions allowing both supportive and opposition sensors, our separate models for Trump and Clinton poll support identified sensors for Hillary Clinton who were disproportionately women and for Donald Trump who were disproportionately white. The method did not predict changes in levels of undecideds and underestimated support for Donald Trump in September 2016, where the errors were correlated with discussions of protests of police shootings.

Campaign observers monitor election polls over the course of campaigns to better understand the factors that determine support for candidates. After the election, analysts can look back at the events and statements that changed political support to help explain reasons behind votes. Although poll movements often do not help predict the election outcome, especially early in campaigns (Gelman and King 1993), the poll movements do convey information on enthusiasm (Gelman et al. 2016) and vote choice (Hillygus and Jackman 2003).

Election and public opinion models typically aggregate many polls that take into account the opinions of large samples of people. These principled methods can be based on representative samples or samples later explicitly weighted to resemble a representative sample (Wang et al. 2015). More speculative social media based poll models usually aggregate many distinct signals of election related sentiment and discussion across a large number of unrepresentative people (Tumasjan et al. 2010; Ceron et al. 2014; Beauchamp 2016). A recent, very successful prediction effort, for example, developed a method based on aggregate word use on Twitter to predict state poll numbers (Beauchamp 2016). Last, some offline works take a prediction market like approach, asking individuals not how they personally feel about the candidates but who they think will win an election (Rothschild and Wolfers 2013) or running tournaments to identify people who are very good at estimating how likely a geopolitical event is to occur (Mellers et al. 2015).

Here, we draw on ideas from polls, prediction markets, and social media based poll estimates. We want to know 1) whether small numbers of regular people's activity on Twitter can accurately track the polls (do we need 1,000, 100, or 10 people to do it?), and 2) if we can identify small numbers of people whose activity levels track the polls, are they different from the general population (in partisan affiliations or demographic characteristics)?

We view these individuals as *social sensors*: individuals whose Twitter streams present a sizable and reliable set of signals about the state of the campaign, correlated with movements of survey respondents. A good *set of* sensors, similarly, provides relatively non-overlapping signals, reflecting the movement of diverse voters.

We describe a straightforward method to select a good set of sensors, describe the characteristics of the selected individuals, and evaluate the extent to which a small (< 20) number of people can predict the polls. Perhaps more important than the method itself, we train our models on a set of social media users whose accounts were linked to voter registration records. This matters because even if the individuals are not representative, we know, with high likelihood, that they at least appear to be members of the American electorate and therefore the signals they provide are less likely to be spuriously associated with the polls in our training period. This is particularly important on Twitter, where an unfiltered sample can contain a large number of bots and anonymous or troll accounts.

Data

We use three data sources for study: Twitter, voter registration records, and public opinion polling averages.

We collected Twitter data over the course of the presidential election campaign for a 22,853 person sample of Twitter

users previously matched to a national sample of voter registration records. The matching used Twitter accounts that list a U.S. location, and it identified $(Twitter\ user, voter)$ pairs for which the full name was unique within their state in both data sets. For Twitter users that we were able to match to voter registration records, we collected tweets that referred to Clinton or Trump. The tweets were identified using a keyword list of variants on the candidates' names. To reduce false positives (such as "Clinton, New Jersey"), we trained a bag-of-words classifier to identify election-related posts and applied it as a filter to this collection. Approximately 5,000 people in this sample mentioned either candidate (either in plain text, "Hillary Clinton", or in handles, "@realDonaldTrump") between May 1, 2016 and November 8, 2016.

For this voter sample, we obtained basic demographic information from their registration records, including year of birth, gender, party affiliation, and race/ethnicity.

After filtering the sample to include only tweets that mentioned the candidates (either their handles or simply their names), the data entered our model as two variables for each account, logged counts of Hillary Clinton mentions and logged counts of Donald Trump mentions, per day of the election campaign May 1 through November 8, 2016. We removed accounts with lower than the sample median activity.

The polling averages were taken from the HuffPost Pollster API [1]. In these polling averages, we create models for poll numbers of 1) Hillary Clinton 2) Donald Trump, and 3) undecided or any other candidate. Our focus is the poll margin, the difference between Hillary Clinton's and Donald Trump's poll numbers.

Methods

Our goal is to identify a small set of sensors who provide non-redundant signals that affect the polls. Ideally, these sensors would either reflect a diverse American electorate or provide diverse signals to approximate it.

One way to select such sensors is to analyze their social networks and news exposure and select active users from clusters of social media accounts. Here, we simply leverage correlations between tweet activity and the polls.

The intent is in the spirit of methods that identify latent relationships using correlated errors in large time series cross section data (Bai 2009). One way to develop such a model with Twitter data is to conduct a principal component analysis on all tweet activity time series and predict the polls using a penalized regression on the top principal components. This approach mirrors methods for creating synthetic controls (Xu 2016). However, given homophily, polarization in news consumption, and very high levels of activity and information aggregation among some Twitter users, we wanted to know whether it was possible to skip the latent relationship measurement step and simply use the activity of a select group of users as sensors for those low-dimensional relationships.

[1] http://elections.huffingtonpost.com/pollster/api

We specifically assessed whether a regression with $l1$ norm regularization (LASSO) on social media activity would give good predictions and provide sensors that were diverse and in touch with non-overlapping sectors of the electorate.

Like an ordinary linear regression, the LASSO minimizes the sum of squared residuals but adds a constraint that shrinks the coefficient estimates based on the absolute value of the coefficients. With this regularization, many of the coefficient estimates shrink to zero, especially in the case where the number of variables greatly exceeds the number of observations ($p >> n$). The maximum number of non-zero coefficients is n (Efron et al. 2004).

In the case of correlated variables, as occurs in our data that uses a large number of similar variables drawn from Twitter users, the LASSO will randomly select a single variable from among many correlated variables (Efron et al. 2004; Zou and Hastie 2005). This is often a drawback of the method and an advantage of the Elastic Net (Zou and Hastie 2005), which can select more than n variables. In our case, however, where each of the variables measures the same quantity, this behavior potentially leads to representative variable selection.

The LASSO selects a small number sensors to track the poll numbers using model fitting in a training period. To create a *distribution* of potential sensors and their demographic characteristics (used in characterizing the sensors rather than for predictions), we implement a simple procedure that reruns the regression without half of the accounts previously selected by the LASSO. Specifically, we run a LASSO, save the user names with non-zero coefficients, and re-run the model without half of those users. We then again store the user names with non-zero coefficients and re-run the model without half of the users from both the first and the second run of the model. This procedure tends to select new accounts for the new runs, especially accounts that were not selected in prior runs due to high collinearity with the top sensors.[2] We run the models 100 times to obtain a distribution of sensors.

We use the number of times that an account is selected for the predictive models as a variable of interest in our assessments of the demographic diversity of the sensors for each of the presidential candidates. We assess the selection with only a coarse measure of variability, however. An extension of estimators to calculate the statistical significance of clusters of variables, such as the joint significance of all accounts of black users compared to white users, is beyond the scope of this paper.

Results

Because we ran separate models for each candidate, we obtained different sensors and numbers of sensors for each model. The models selected 6 sensors for Clinton and 14 for Trump. One sensor was included in both of the models. There were 25 sensors in the undecided/other candidate model, 5 of whom were in the Clinton or Trump models.

[2] The total number of accounts selected by a large number of runs resembles the variable selection of an Elastic Net.

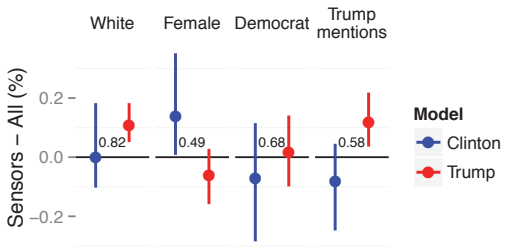

Figure 1: *Sensor description, median and interquartile range.* For this figure, we calculate the mean of the demographic characteristics in each re-run of the model training and show the median of the runs, along with the interquartile of the mean characteristics in the runs.

Figure 1 shows the distribution of demographic characteristics among potential sensors using the rough estimate of variance described in the methods section. The y-axis is the difference between the proportion of sensors for each candidate with a particular characteristic (e.g. race) (weighted by the number of times an account was included in the models) compared to the sample average (weighted by the number of political tweets). The text above the black lines at 0.0 on the y-axis state the weighted average of a characteristic in the original sample.

The figure shows that Clinton's sensors were disproportionately female, while Trump's sensors were disproportionately white or male. The heavily white Trump sensors and the relatively female Clinton sensors are surprising given that the models allow both positive and negative sensors, and we do not find that the sensors are heavily co-partisan.

In Figure 2, we show predictions based on the sensors' candidate related activity compared to the poll average. The left panel shows the poll margins for Hillary Clinton against Donald Trump and the right panel shows the overall undecided or support for third party candidates responses. In the left panel, the mean absolute error of the predictions in the test period was 0.69. By comparison, the mean absolute error of the poll mean in the test period was 1.37.

Like the national polls, the predictions detect a decline in support for Hillary Clinton compared to Donald Trump that reaches its lowest point around September 11, 2016 (either on or a few days before Hillary Clinton fainted during a September 11 memorial service). It further detects the jump in support for Hillary Clinton against Donald Trump during the presidential debates September 26 through October 19, peaking around October 7 (release of Billy Bush/Donald Trump video, along with Wikileaks Podesta email release).

We assess whether these users are simply following and reporting the polls rather than following campaign events and rhetoric by counting the number of times the sensors mention the polls compared to occurrence in the overall sample. Of the tweets in the original sample, 1.6% of the tweets used the word "poll" and of the tweets used by the sensors, 1.7% of the Clinton sensors' tweets and 1.8% of the Trump sensors' tweets used the word "poll". Meaning, the sensors did not mention the polls at unusually high levels.

After taking the second order difference of both the poll margins and predictions (the number of differences necessary here to reject a unit root at 95% significance using the Augmented Dickey-Fuller test), the predictions Granger-caused (i.e. led) the polls (p-value 0.06) and the polls also Granger-caused the predictions (p-value 0.13). In other words, we find some evidence (n=84) that the sensors predict the polls and that they are simultaneously influenced by them.

Discussion

The findings show that the activity of very small numbers of individuals can predict certain opinions for an entire country. Unlike polling or population based sampling methods based on online samples (Wang et al. 2015), we did not select or weight these individuals to be a representative sample of the entire electorate. The method was also not closely related to prediction market methods, where sensors assess information around an election and attempt to predict the outcome (Rothschild and Wolfers 2013). Instead, the sensors tracked and predicted the polls unintentionally through their online social media discussion of the candidates. This resembles other machine learning based methods for predicting polls using social media data (Tumasjan et al. 2010; Ceron et al. 2014; Beauchamp 2016), but did not use any sentiment detection and did not aggregate word usage across individuals.

Important to future work in detecting sensors using more principled and social network based methods, we illustrate that the model selects sensors in theoretically interesting and potentially important ways. Somewhat surprisingly, the sensors for Hillary Clinton were disproportionately female, while the sensors for Donald Trump were disproportionately white – even though the samples were not starkly different on partisanship. This could reflect a sharp disconnect between the considerations that were salient to supporters and detractors of the candidates (and that did not fall along the party lines).

A sensor method appears to hold promise, but will require more work to make it more robust and generalizable. We do not think that we could use the same sensors for a wide variety of contexts. Many of the users mostly tweet about politics. Perhaps due to the sensors' high levels of political interest, we were not readily able to identify sensors who tracked the undecided/other candidate responses in the polls for both the pre and post August 15 periods. The method is also heavily reliant on the amount of movement present in the polls. Here, the method is limited by the very small number of sharp shifts in poll numbers before August 15.

We used our prediction errors to (speculatively) identify political events that were not accurately reflected in our poll predictions. To do this, we read through news reports in late August and September to identify campaign events that could have affected the polls. We narrowed the set of important events to two: 1) protests of police shootings, including National Anthem protests and the Charlotte riots, and 2) discussion of the health of the candidates. Based on Twitter data

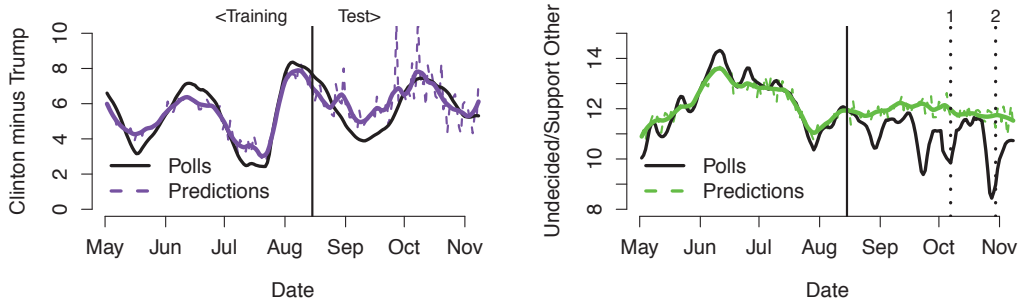

Figure 2: *Poll predictions from voter sensors.* The left panel of this figure shows the poll margin predictions using the difference between the sensor model predictions for Clinton's support and Trump's support. The right panel of this figure shows the predictions for an equivalent model where the outcome is undecided/other candidate in the same poll average. We used loess smoothing so that our predictions average nearby predictions much like the poll smoothing on which our models are based.

from Crimson Hexagon [3], discussion of Hillary Clinton's and Donald Trump's health spiked only around September 11 and so was not correlated with the prediction errors. The prediction errors were, however, correlated with the negative sentiment discussion of protests of police shootings (Figure 3). These prediction errors were primarily driven by an underestimation of Donald Trump's support rather than an overestimation of Hillary Clinton's.

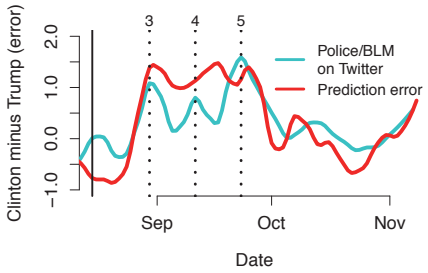

Figure 3: *Comparison of prediction errors and discussions of protests of police-involved shootings.* This figure shows the logged, scaled, and smooth counts of negative sentiment tweets containing "police shooting", "law and order", "protest", "riot", "black lives matter", and "BLM" in the United States compared to our prediction errors.

Acknowledgements

This work was generously supported by the National Science Foundation (NSF) CDI-Type II grant no. 1125095 and Army Research Office (ARO) grant no. W911NF-12-1-0556.

References

Bai, J. 2009. Panel Data Models With Interactive Fixed Effects. *Econometrica* 77(4):1229–1279.

Beauchamp, N. 2016. Predicting and Interpolating State-level Polls using Twitter Textual Data. *American Journal of Political Science* 1–36.

Ceron, A.; Curini, L.; Iacus, S. M.; and Porro, G. 2014. Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. *New Media & Society* 16(2):340–358.

Efron, B.; Hastie, T.; Johnstone, I.; and Tibshirani, R. 2004. Least Angle Regression . *The Annals of Statistics* 32(2):407–499.

Gelman, A., and King, G. 1993. Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable? *British Journal of Political Science* 23:409–451.

Gelman, A.; Goel, S.; Rivers, D.; and Rothschild, D. 2016. The mythical swing voter. *Quarterly Journal of Political Science* 11:103–130.

Hillygus, D. S., and Jackman, S. 2003. Voter Decision Making in Election 2000: Campaign Effects, Partisan Activation, and the Clinton Legacy. *American Journal of Political Science* 47(4):583–596.

Mellers, B.; Stone, E.; Murray, T.; Minster, A.; Rohrbaugh, N.; Bishop, M.; Chen, E.; Baker, J.; Hou, Y.; Horowitz, M.; Ungar, L.; and Tetlock, P. 2015. Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science* 10(3):267–281.

Rothschild, D., and Wolfers, J. 2013. Forecasting elections: Voter intentions versus expectations. 2011. *SSRN*.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*.

Wang, W.; Rothschild, D.; Goel, S.; and Gelman, A. 2015. Forecasting elections with non-representative polls. *International Journal of Forecasting* 31(3):980–991.

Xu, Y. 2016. Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models. *Political Analysis*.

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

[3] https://forsight.crimsonhexagon.com/