Converting the Twitter API into an Online Panel via Human Intelligence to Measure Public Opinion

Roberto Cerina
roberto.cerina@nuffield.ox.ac.uk
Nuffield College, University of Oxford

Raymond Duch
raymond.duch@nuffield.ox.ac.uk
Nuffield College, University of Oxford

November 3, 2020

1 Summary

We use Human Intelligence provided by Mechanical Turks to convert the Twitter streaming API into a structured online panel, to monitor opinion over the 2020 election campaign in the US. We leverage opinion about the horse-race of both Twitter users and Turks, as they can be obtained simultaneously with a single survey instrument, to produce our estimates. A Random Forest is trained on the resulting data, and predictions from this learner are stratified according to a stratification frame made up of likely voters, provided to us by `Øptimus Analytics`[1]. Results are aggregated up to the state level to produce state-level estimates of opinion. A global correction factor to account for online-selection is calculated by comparison with publicly available data and added to the estimates. The results are close (under 4% MAE overall, and under 3% for the most competitive states) to FiveThirtyEight's election-day forecasts, though they do present challenges especially in the least competitive states.

2 Background

Polling public Opinion via Random Digit Dial (RDD) is the go-to method for high-quality polling in the US; this is evidenced by its dominance in the FiveThirtyEight Pollster Ratings[27]; its performance in comparison to internet surveys[31] as well as its use by large media corporations as the go-to example to show how polling works to the wider public[29]. It is therefore worrisome that response rates have plummeted below 6%[16, 14]. Though recent reviews of historical polling accuracy have shown little evidence for the decrease in accuracy of national polls overall[26, 13], the polling report on the 2016 US election[15] highlighted important over-sampling of college graduates in mid-western states - a factor that led polling aggregators relying on these polls to assign inaccurate probabilities to the realization of a Trump victory. The report makes clear that adjusting via re-weighting the sample to match key characteristics in the population is necessary to recover correct estimates of the vote. Given this necessity to use model-based adjustments, pollsters in others countries - in particular the UK - have started moving

[1] `https://0ptimus.com/`

away from expensive and labour-consuming RDD[2], and toward online panels of internet-survey respondents[19, 11]. Multilevel Regression and Post-Stratification (MrP)[8, 7] is then used to make representative small-area inference from these unrepresentative samples. Though these online panels have shown promise, and have been shown to accurately reflect public opinion provided appropriate modeling adjustments, a recent study by PEW[17] shows that as many as 7% of these online responses could be bogus, i.e. not reflecting the respondents' true sentiments, and that bogus responses are not random - but rather systematically biased.

On the spectrum of *non-probability*, online panels are less extreme than social media, as evidenced by recent reviews of the evolving demographics of Twitter and Facebook[23]. Nevertheless, at the individual-level, Barberá[1] has shown we can accurately predict one's political preferences via their network of friends on Twitter. This suggests that, if we were able to accurately re-weight these individuals according to their true probability of inclusion in the population, we could recover accurate measures of political opinion. The Twitter streaming API is free to use for research purposes and can allow for harnessing several millions of Tweets from thousands of users with ease from one's laptop. It promises to render a real-time random sample of the Tweets which contain certain search-terms, though some have shown these samples are unlikely to be random in practice[4]. Being able to extract signal from this source opens the door for real-time, unobtrusive measurement of public opinion on a massive scale.

3 Methodology

The challenge we face is that, in order to post-stratify our estimates, we need to be able to reconcile the user-data we obtain via Twitter with that available in a stratification frame, which are typically derived from the census or from voter-registration files in the US. Hence we need to be able to restructure the information available on Twitter: the variable 'gender' for instance will be dispersed in unorganized fashion between the profile-picture, the description, the tweets, and within other information about a given user provided to us by the twitter API. We use Human intelligence, provided by Amazon Mechanical Turk workers, to perform this task. We can think of this as a 'wisdom of crowds'[24] approach to extracting survey responses from Twitter: provided each worker has a probability larger than 0.5 to correctly label a given twitter account, we can be confident that as the number of workers increases we recover accurate information about a given Twitter users - as if they had taken a survey themselves.

In this paper we use a single survey instrument to accomplish 2 tasks: i) we display information about Twitter users to Mechanical Turk workers for them to label; ii) we ask Workers about themselves. We use signal from both sources to inform a model aiming to estimate public opinion on Election-day in the US. Note that Mechanical Turks are themselves a highly selected population[12, 5], though past work has shown reasonable potential for representative inference provided model-adjustments are implemented[9]. We implement attention-checks to control data-quality, as well as to screen-out bots and VPN workers falsifying their location[18]. Twitter Bots were excluded to the extent that Mechanical Turk workers could recognize an account as *'Definetly a Bot'*, the highest score in a five items Likert scale question.

3.1 Data Collection

We began collecting Twitter data on August $1st$ 2020, and completed our collection on November 3^{rd}. Around 3.5 million tweets containing the terms 'Biden' or 'Trump' were obtained over the course of the campaign. From these, we screened-out non-english tweets to limit the amount of

[2] Typical costs for live-call RDD polling per respondent can range from 5 to 15 USD, depending on the population which is being surveyed. In comparison, online panel respondents can be recruited for around 1 USD.

non-US users in our pool; we further screen out all users who tweeted less than 10 times during the period of study. At the end of the monitoring period, this group accounted for just under $35,000$ users, responsible for around 1 million tweets. During the campaign, we sampled from this group between 500 and 1000 users, and stored their information into a database which was then rendered to Mechanical Turk workers for analysis, on the same day. Over the course of the study we obtained labels for $5,350$ Twitter users, and $2,500$ unique Mechanical Turk workers. Each user and/or workers could contribute up to once per day, hence the total number of traces amounts to just under $10,000$, after accounting for workers which left multiple traces over the course of the monitoring period. Each worker responding was paid 1.25 USD to complete a survey, and from each survey we obtained 2 traces - one for the Twitter user and one for the Turk; hence the theoretical cost-per-trace with this approach is 0.625 USD. After accounting for responses we threw away due to data-quality issues evidenced by attention- and bot- checks, the cost per trace was around 1 USD.

3.2 Covariate Space

With respect to the substantive variables we collected for each Turk and Twitter user, we collected the following: {*Days to Election; State; Region; Gender; Ethnicity; Yearly Household Income Bracket; Turnout 2016; Turnout 2018; Party ID; Marital Status; Age Bracket; College Degree Ownership; 2016 Vote Choice; 2020 Vote Choice*}. These variables were chosen upon receiving our stratification frame from `Øptimus Analytics`. The frame is based on registered voters according to the L2 voter registration files; the most updated frame at our disposal is dated October 29^{th}. An example of the frame follows in Table 1.

To the individual-level variables, we added a series of area-level covariates[3]: { *2016 R Pres. Vote; 2016 R-D Pres. Margin; 2012 R Pres. Vote; 2012 R-D Pres. Margin; 2020 Senate Incumbency; 2020 House Incumbency; % Pop. 17 to 25 ; % Pop. 26 to 34; % Pop. 35 to 54"; % Pop. 55 to 64"; % Pop. 65+; % Pop. Hispanic; % Pop. Black; % Pop. Asian; % Pop. White; % Females above 15 who never married; % Foreign Born; Population Size; Land Area km^2; Population Density; Area-Weighted Population Density; % Working in Management/Business/Science/Arts; % Working in Primary Sector/Construction/Maintenance; % Working in Manufacturing; Average Household Size; Median Household Income; % in Poverty; % College Degree; % Non-College White; % College Non-White; % Civilian Veterans; Number Evangelical or Mormons; Number Catholics*}.

We finally added some variables which varied through time and space, along with some interactions: {*Cumulative COVID Cases Count; Cumulative COVID Deaths Count; 2020 GDP Growth (Quartely); 2020 Unemployment Rate (Monthly)*}.

[3] Several of these were derived from the Daily Kos Data repository - `https://www.dailykos.com/stories/2018/2/21/1742660/-The-ultimate-Daily-Kos-Elections-guide-to-all-of-our-data-sets`

state	t20	gender	ethnicity	hh_income	t16	t18	party_code	marital_status_code	age_bins	modeled_college_grad	vote2016	count
AK	0.00	M	European	[75000, 100000)	0.00	0.00	Nonpartisan or Other	Not Married	35-44	Modeled No College Degree	Modeled Voted Clinton	91.00
AK	0.00	M	European	[75000, 100000)	0.00	0.00	Nonpartisan or Other	Not Married	25-34	Modeled No College Degree	Modeled Voted Clinton	76.00
AK	0.00	M	European	[75000, 100000)	0.00	0.00	Nonpartisan or Other	Not Married	35-44	Modeled No College Degree	Modeled Voted Trump	74.00
AK	0.00	F	European	[75000, 100000)	0.00	0.00	Nonpartisan or Other	Not Married	35-44	Modeled No College Degree	Modeled Voted Clinton	73.00
AK	0.00	F	European	[100000, max]	0.00	0.00	Nonpartisan or Other	Not Married	35-44	Modeled No College Degree	Modeled Voted Clinton	61.00
WY	1.00	F	European	[100000, max]	0.00	1.00	Nonpartisan or Other	Not Married	18-24	Modeled No College Degree	Modeled Voted Trump	1.00
WY	1.00	F	European	[75000, 100000)	0.00	0.00	Republican	Not Married	55-64	Modeled College Degree	Modeled Voted Trump	1.00
WY	1.00	F	East and South Asian	[50000, 75000)	1.00	1.00	Republican	Not Married	55-64	Modeled No College Degree	Modeled Voted Trump	1.00
WY	1.00	M	European	[min, 25000)	1.00	1.00	Republican	Not Married	25-34	Modeled College Degree	Modeled Voted Trump	1.00
WY	1.00	F	Other	[100000, max]	1.00	1.00	Democrat	Not Married	45-54	Modeled College Degree	Modeled Voted Clinton	1.00

Table 1: A selection of rows from the stratification frame we are using. The ‘turnout 2020’ variable is produced according to a proprietary model by Øptimus Analytics, and is informed by a statistical model incorporating historical and demographic data from voter registration files.

3.3 Turnout

With respect to turnout, we relied on reported turnout for our sample: a five items Likert scale was used to detect turnout propensity; workers self-reported their turnout likelihood, and predicted the likelihood of turnout of the Twitter users based on the information available to them. We took the top-two responses to indicate that a user was a likely voter. `Øptimus Analytics` used a proprietary model to classify individuals in the registration files as 'likely voters' - this can be seen in Table 1.

3.4 Learning

We privilege an implementation of the Regression and Post-Stratification approach which uses Random Forests (RF)[3], implemented in `R`[28] via the `ranger`[30] package. The estimation procedures with RF is significantly faster than with Multilevel Regression, and does not require the modeler to perform ad-hoc variable selection - rather, we can rely on the forest to seek an optimal set of covariates. Empirically, we noticed that pre-specifying interactions between variables, and including these amongst the set of covariates to be considered, makes it more likely that these be taken into consideration. Due to the significant importance assigned to individual-area level interactions in the 2019 UK election MrP forecasts[19, 20, 10] we augment the covariate space with thousands of individual-area interactions. We also include measures of physical distance across states to allow the model to take local smoothing into consideration.

The hyper-parameter specification of our forests is justified as follows: `num.trees` is set to $1,000$, a large-enough number to avoid monte-carlo effects and search through a large number of different trees to extract the most meaningful signal; `mtry`, the number of variables to consider splitting in each tree, is set to P/3, where P is the total number of covariates. This is a choice larger than the default $\sqrt{P}$, motivated by findings showing the inclusion of a larger than normal number of variables is necessary in probability estimation via RFs[25]; `min.node.size` is set to 10% of the sample size, as per Malley et al.[22]; `sample.fraction` is set to 2/3, and `replace = FALSE`, to limit the potential similarity of each tree, and hence enhancing the error-reduction potential of aggregating the weak learners.

3.5 Adjusting for Online Selection

There remains a degree of online selection which will require to adjustment beyond the variables we have in our stratification frame. There is no doubt that selecting into Twitter or the Mechanical Turk platform will be residually correlated with vote-choice, and more thinking has to go into identifying the exact channels to see if they can be accounted for at the individual level. In absence of that, we simply implement an intercept shift to our estimates; this is estimated as the average difference between our predictions and a polling average - in our case, the FiveThirtyEight election-day predictions. We find the intercept shirt to be worth 4 percentage points in the direction of Trump.

4 Sample v. Population Comparison

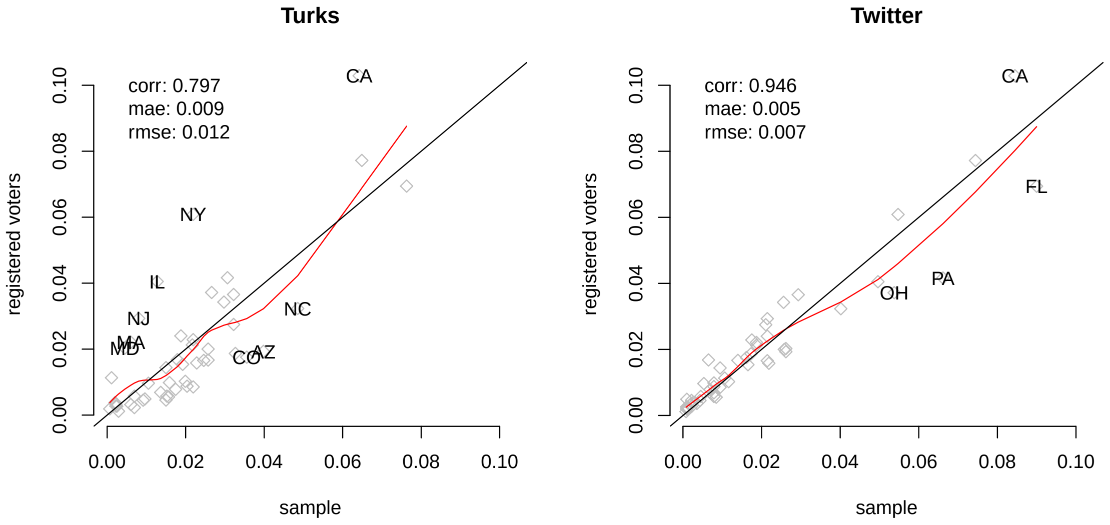

Figure 1: Sample v. Population Comparison: States.

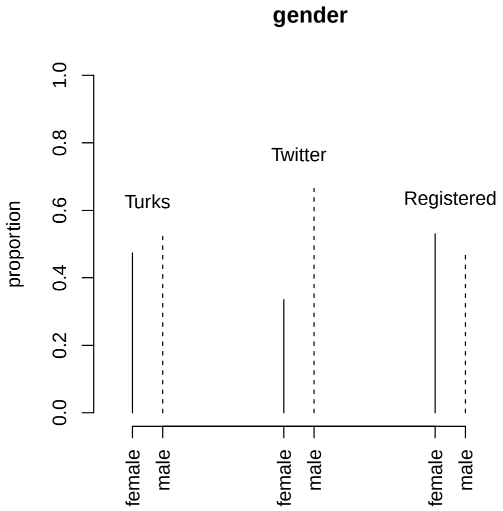

Figure 2: Sample v. Population Comparison: Gender.

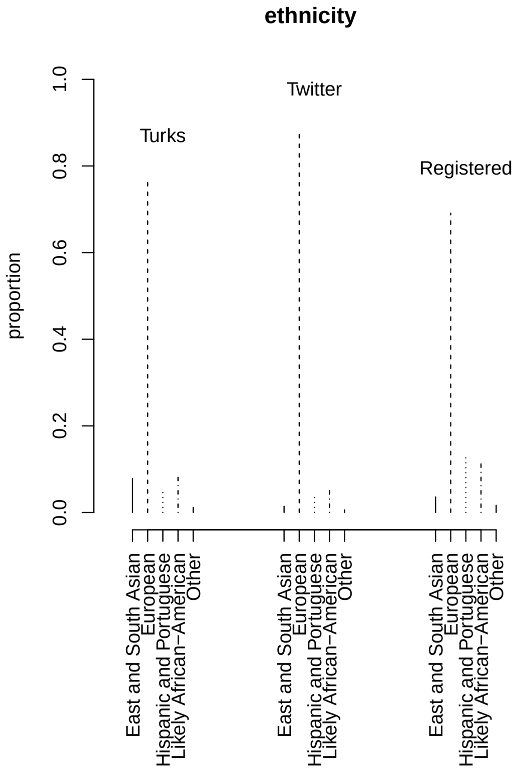

Figure 3: Sample v. Population Comparison: Ethnicity.

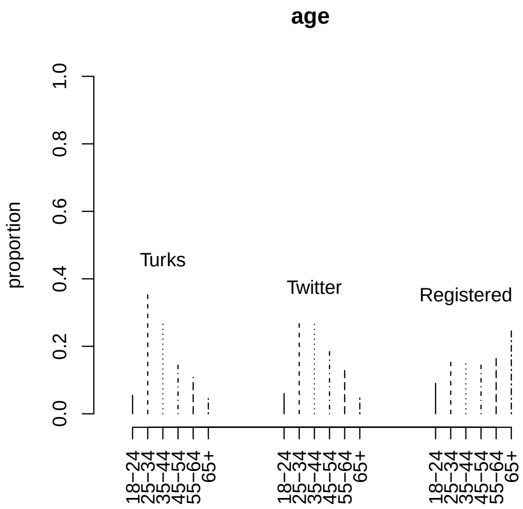

Figure 4: Sample v. Population Comparison: Age.

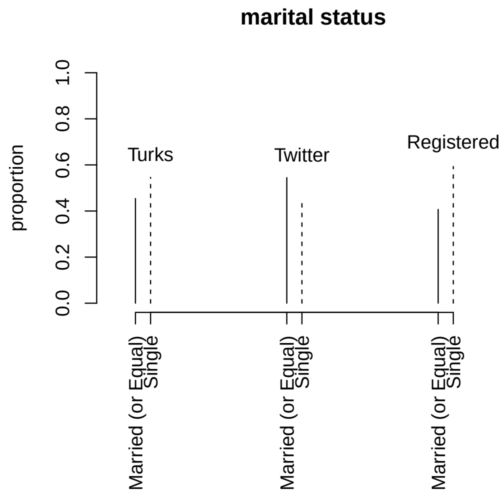

Figure 5: Sample v. Population Comparison: Marital Status.

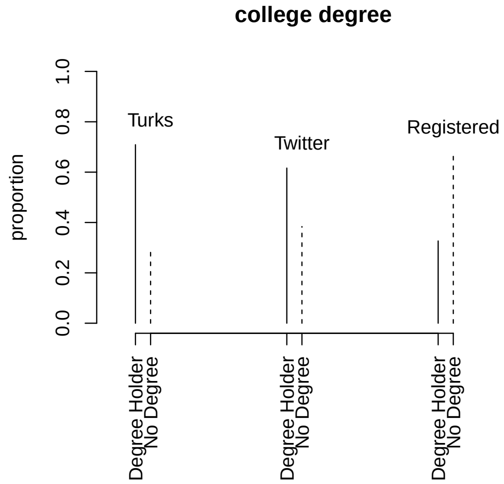

Figure 6: Sample v. Population Comparison: College Degree.

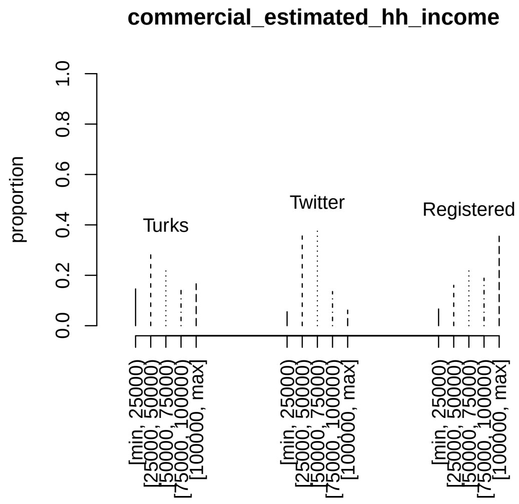

Figure 7: Sample v. Population Comparison: Household Income.

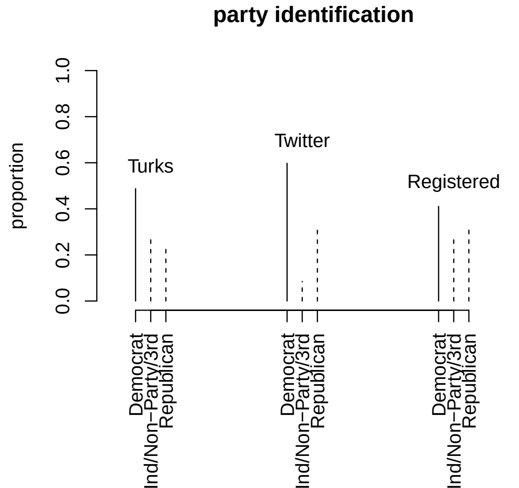

Figure 8: Sample v. Population Comparison: Party Identification.

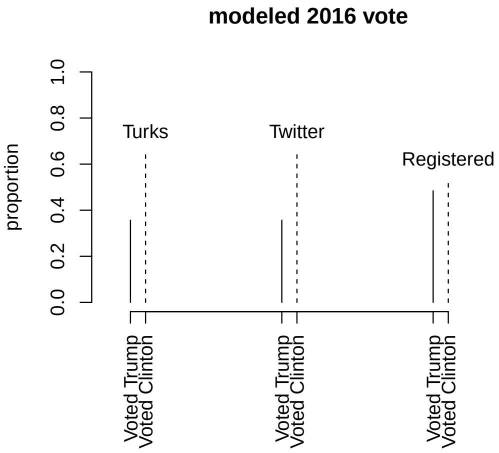

Figure 9: Sample v. Population Comparison: 2016 Vote Choice.

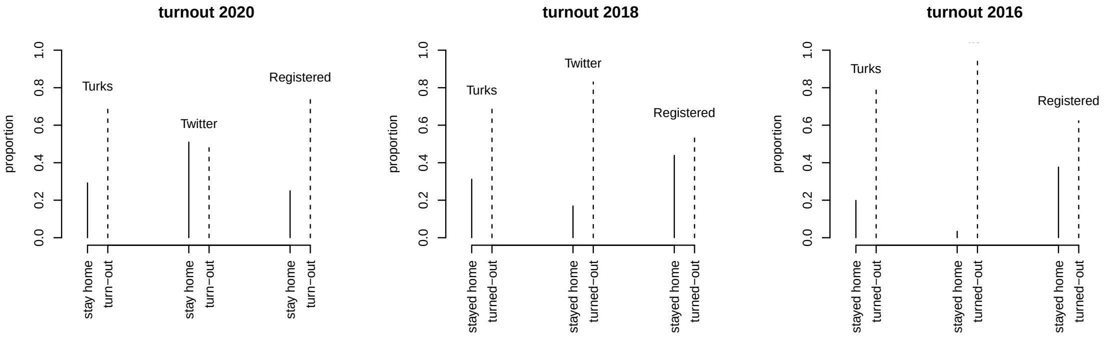

Figure 10: Sample v. Population Comparison: Turnout over elections.

5 Area-Level Estimates

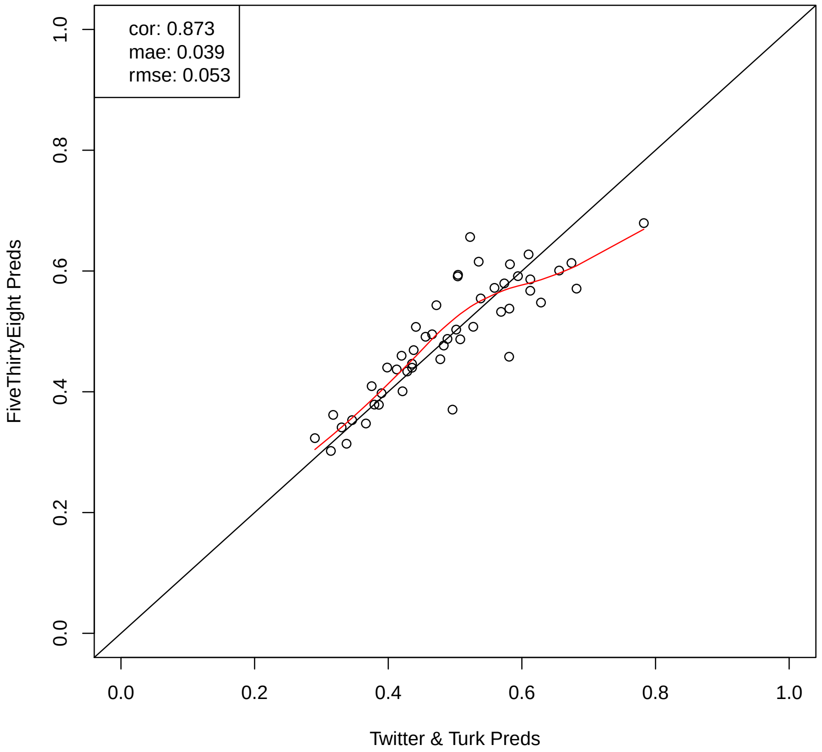

Figure 11: Comparison between Twitter + Turk forecast and FiveThirtyEight's election-day predictions.

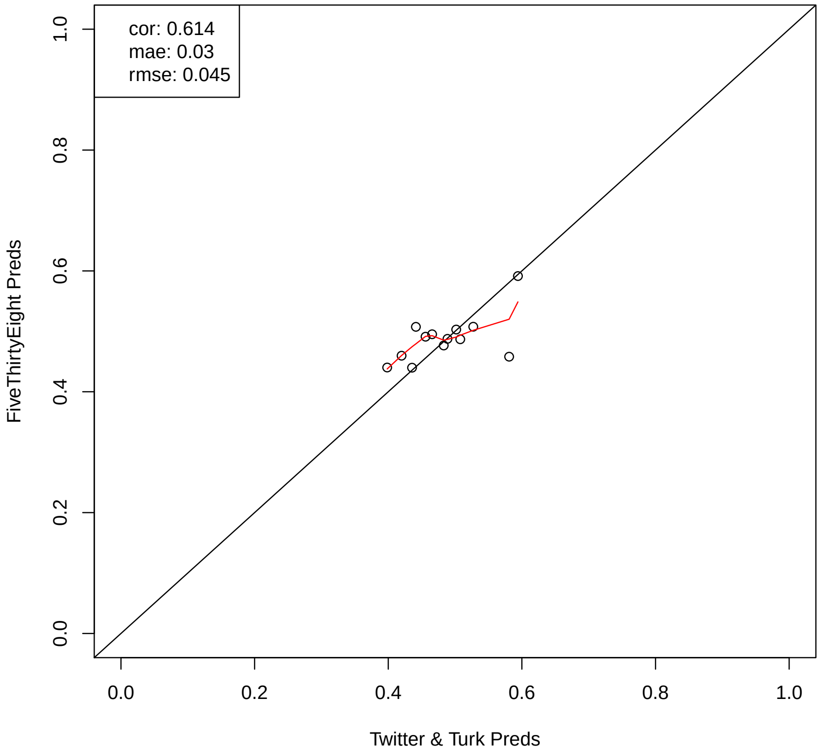

Figure 12: Comparison between Twitter + Turk forecast and FiveThirtyEight's election-day predictions - swing-states only. The outlier is Wisconsin

6 Discussion & Future Work

Future work we plan to undertake with this data, and which we think will make our modeling effort more successful, includes the accounting for Turk-characteristics in the labeling exercise - specifically with respect to the vote-choice and turnout questions: it is possible that partisan Turks' predictions about partisan Twitter users' are systematically biased; if we could net this bias prior to modeling voting intention, we could obtain better estimates of the vote, particularly considering the severe partisan split amongst mechanical Turks.

A further area of research we plan to undertake with this data is the analysis of individuals who have changed their mind over the course of the campaign. These are relatively few in our sample - numbering in the hundreds - but can still provide an interesting test to evaluate the extent to which Human Intelligence can capture the evolving preferences of individuals.

Something which we have not yet implemented - but which we have a plan to address - is dealing with uncertainty for RF predictions. We plan to leverage the `forestError` package by Lu and Hardin[21], which leverages out-of-bag samples constructed as a by-product of forest-building to estimate Mean Squared Prediction Error (MSPE) of the forest.

Finally we would like to evaluate the extent to which the learner choice is contributing to the error of our predictions. A direct comparison between RFs and Multilevel Regression will reveal if the speed-gains provided by the forest come at unacceptable accuracy costs. A-priori, we don't believe this should be the case, given the good performance of RFs in a variety of settings[6], and it would be strange if Opinion Polling was the lone application to which this algorithm cannot be applied. Moreover, other tree-based learners have shown superiority in these direct comparisons[2], and the recent expansion of MrP models to include vast swaths

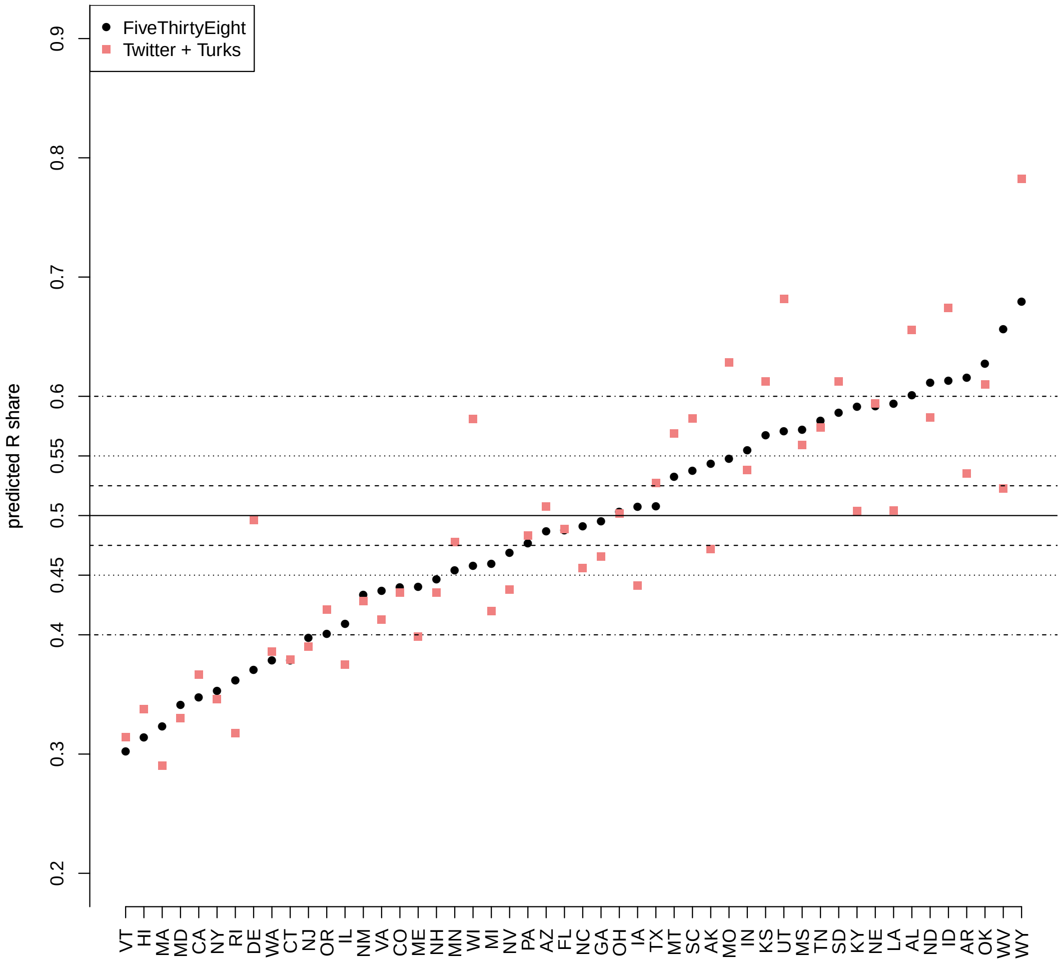

Figure 13: Comparing state-level predictions of our model (red) against FivetThirtyEight (black).

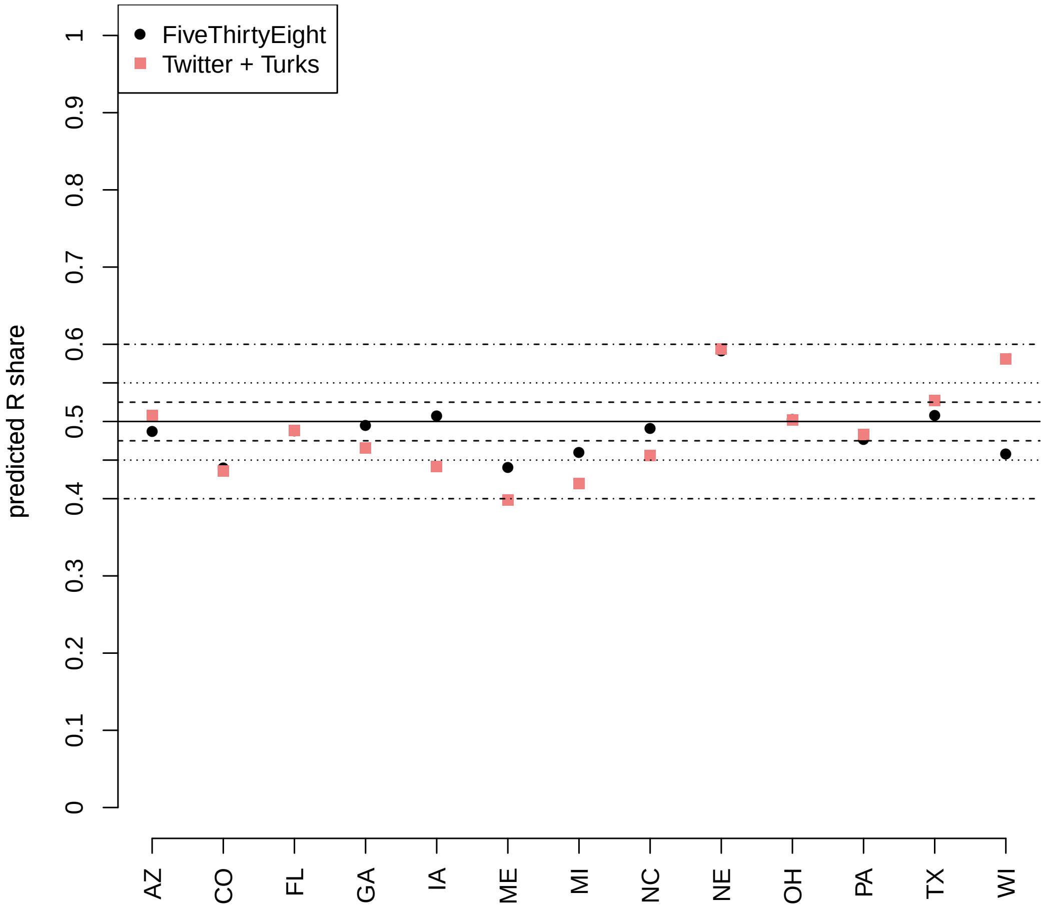

Figure 14: Comparing state-level predictions of our model (red) against FivetThirtyEight (black) - swing-states only.

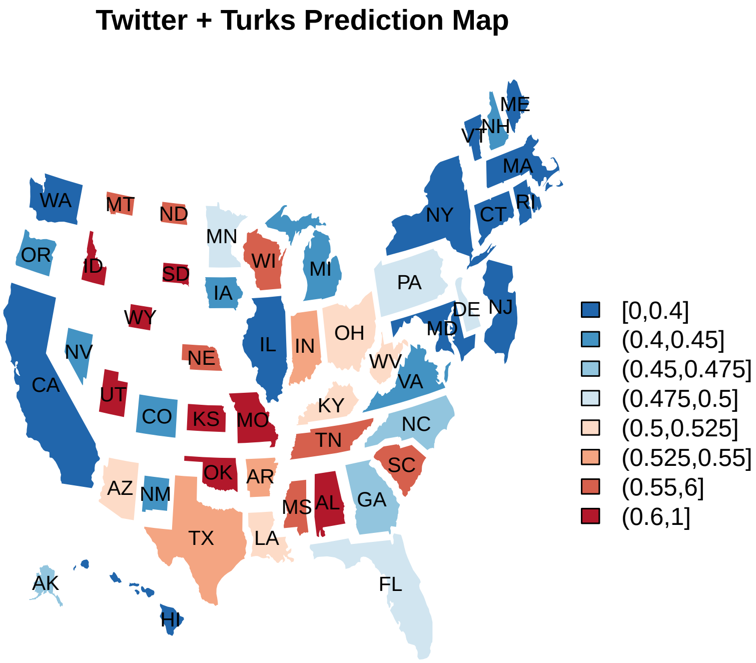

Figure 15: A map showing which states will break for whom. Assuming Nebraska 2nd, Maine 2nd and DC all go to Biden, as is the case in the FiveThirtyEight predictions, the above map suggests Biden will be elected president, with an Electoral Vote tally of 339 against a Trump tally of 199.

of individual-area interactions suggest non-linearities are at play. Others however have argued there aren't enough non-linearities in vote-choice models, and that polling samples are too small for the RFs to powerfully explore full covariate space.

Finally, and most interestingly for the long-run prospects of this project, the data we have gathered in this exercise can be used to train machines to extract survey-like responses from Twitter in real-time. This can allow for real-time monitoring of public opinion at the deep-area level, at no cost to the researcher; the creation of this AI is, in our opinion, the new frontier of opinion polling research.

References

[1] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political analysis*, 23(1):76–91, 2015.

[2] James Bisbee. Barp: Improving mister p using bayesian additive regression trees. *American Political Science Review*, 113(4):1060–1065, 2019.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] Alina Campan, Tobel Atnafu, Traian Marius Truta, and Joseph Nolan. Is data collection through twitter streaming api useful for academic research? In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3638–3643. IEEE, 2018.

[5] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

[6] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

[7] Andrew Gelman. Regularized prediction and poststratification (the generalization of mister p). *URL: https://statmodeling. stat. columbia. edu/2018/05/19/regularized-predictionpoststratification-generalization-mister-p*, 2018.

[8] Andrew Gelman and Thomas C Little. Poststratification into many categories using hierarchical logistic regression. 1997.

[9] Sharad Goel, Adam Obeng, and David Rothschild. Non-representative surveys: Fast, cheap, and mostly accurate. In *Working Paper*. 2015.

[10] Chris Hanretty. Areal interpolation and the uk's referendum on eu membership. *Journal of Elections, Public Opinion and Parties*, 27(4):466–483, 2017.

[11] Chris Hanretty. 2019 General Election MRP predictions. `https://www.survation.com/2019-general-election-mrp-predictions-survation-and-dr-chris-hanretty/`, 2019.

[12] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Benjamin V Hanrahan, Jeffrey P Bigham, and Chris Callison-Burch. Worker demographics and earnings on amazon mechanical turk: An exploratory analysis. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.

[13] Will Jennings and Christopher Wlezien. Election polling errors across time and space. *Nature Human Behaviour*, 2(4):276–283, 2018.

[14] Scott Keeter, Nick Hatley, Courtney Kennedy, and Arnold Lau. What low response rates mean for telephone surveys. *Pew Research Center*, 15:1–39, 2017.

[15] Courtney Kennedy, Mark Blumenthal, Scott Clement, Joshua D Clinton, Claire Durand, Charles Franklin, Kyley McGeeney, Lee Miringoff, Kristen Olson, Douglas Rivers, et al. An evaluation of the 2016 election polls in the united states. *Public Opinion Quarterly*, 82(1):1–33, 2018.

[16] Courtney Kennedy, Nick Hatley, Andrew Mercer, Arnold Lau, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo. Response rates in telephone surveys have resumed their decline. `https://www.pewresearch.org/fact-tank/2019/02/27/response-rates-in-telephone-surveys-have-resumed-their-decline/`, 2019.

[17] Courtney Kennedy, Nick Hatley, Andrew Mercer, Arnold Lau, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo. Assessing the Risks to Online Polls From Bogus Respondents. `https://www.pewresearch.org/methods/2020/02/18/assessing-the-risks-to-online-polls-from-bogus-respondents/`, 2020.

[18] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Philip D Waggoner, Ryan Jewell, and Nicholas JG Winter. The shape of and solutions to the mturk quality crisis. *Political Science Research and Methods*, pages 1–16, 2018.

[19] Benjamin E Lauderdale. How YouGov's 2019 General Election model works. `https://yougov.co.uk/topics/politics/articles-reports/2019/11/27/how-yougovs-2019-general-election-model-works`, 2019.

[20] Benjamin E Lauderdale and Jack Blumenau. Constructing and assessing seat level estimates. Reading the 2019 Election Polls: Event by the London School of Economics, Department of Methodology, and the British Polling Council, 27/11/2019.

[21] Benjamin Lu and Johanna Hardin. A unified framework for random forest prediction error estimation. *arXiv preprint arXiv:1912.07435*, 2019.

[22] James D Malley, Jochen Kruppa, Abhijit Dasgupta, Karen G Malley, and Andreas Ziegler. Probability machines. *Methods of information in medicine*, 51(01):74–81, 2012.

[23] Jonathan Mellon and Christopher Prosser. Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Research & Politics*, 4(3):2053168017720008, 2017.

[24] Andreas Erwin Murr. "wisdom of crowds"? a decentralised election forecasting model that uses citizens' local expectations. *Electoral Studies*, 30(4):771–783, 2011.

[25] Matthew A Olson and Abraham J Wyner. Making sense of random forest probabilities: a kernel perspective. *arXiv preprint arXiv:1812.05792*, 2018.

[26] Christopher Prosser and Jonathan Mellon. The twilight of the polls? a review of trends in polling accuracy and the causes of polling misses. *Government and Opposition*, 53(4):757–790, 2018.

[27] Nate Silver, Derek Shan, Mary Radcliffe, and Dhrumil Mehta. Pollster Ratings. `https://projects.fivethirtyeight.com/pollster-ratings/`, 2020.

[28] R Core Team et al. R: A language and environment for statistical computing. 2013.

[29] The Upshot. Polling in Real Time: The 2018 Midterm Elections. `https://www.nytimes.com/interactive/2018/upshot/elections-polls.html`, 2020.

[30] Marvin N Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

[31] David S Yeager, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang. Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and non-probability samples. *Public opinion quarterly*, 75(4):709–747, 2011.