Voting with your Tweet: Forecasting congressional elections with social media data[*]

Mark Huberty[†]

March 25, 2012

Abstract

This paper demonstrates that the application of machine learning techniques to the raw text of the Twitter message feed can generate highly accuracy predictive algorithms for district-level Congressional election outcomes. Supervised topic modeling of the feed for the 2010 Congressional election generates 75%+ out-of-sample accuracy used on its own. Improving on that accuracy through the use of a cross-validated ensemble learner like the SuperLearner forecasts election outcomes with 86-92% out-of-sample accuracy. In-sample accuracy rates exceed a range of mainstream forecasts, including Congressional Quarterly and Real Clear Politics. Out-of-sample rates do less well, though they continue to equal CQ. However, some evidence indicates that the trained algorithm has an implicit Republican party bias, possibly due to the feed itself reflecting the historic swing nature of the 2010 election. Furthermore, we find some evidence that, in expectation, the algorithm may not consistently do better than predictions based on party incumbency.

---

[*] Paper prepared for the panel on Research Methods in an Internet Era, International Studies Association Annual Conference, April 2012. First version: February 2011. Thanks to F. Daniel Hidalgo, Jasjeet Sekhon, and participants at the 2011 Society for Political Methodology meeting for helpful comments and feedback. All errors remain my own.

[†] Travers Department of Political Science, University of California, Berkeley. Contact: markhuberty@berkeley.edu.

1 Introduction

Microblogging services like Twitter have seen traffic and interest grow rapidly in the last five years. As of September 2010, Twitter itself reported 175 million registered users, generating 95 million messages per day. Despite persistent speculation that much of this content goes unread, Twitter has attracted significant attention from both private and public sector actors as a measure of social sentiment. It has also played a large role in several notable political events, including the 2010 Iranian and 2011 Egyptian and Tunisian political uprisings.[1]

This paper applies machine learning techniques to the content of Twitter messages to predict United States House of Representatives electoral outcomes at better than 85% accuracy. Prediction accuracy compares favorably with mainstream predictions of election outcomes, though out-of-sample tests suggest several caveats. First, predictions of vote percentage are much less accurate than simple win/loss outcomes. Second, some evidence suggests that the trained algorithms have an implicit Republican bias, owing to the strong Republican swing in the 2010 election. Finally, these findings in no way suggest that Twitter content will behave similarly in subsequent elections. Nevertheless, these estimates may still represent a lower bound for potential accuracy based on unexploited information potential in the twitter feed.

---

[1] Twitter's role in the Iranian events was significant enough for the United States Department of State to intervene to postpone scheduled maintenance downtime. Demonstrators in both the Egyptian and Tunisian events of January-February 2011 used Twitter as a coordination and communications medium. In all three instances, however, subsequent entry of government or police agents into the conversation rendered Twitter less useful. For the Iranian events, see Stone and Cohen (2009). For the Tunisian and Egyptian events, see, for example, Kirkpatrick (2011).

2 The Twitter phenomenon

Recent years have seen significant advances in using the sheer volume of data available from Twitter and other internet-based information sources to predict social and economic trends. While much of this has been driven by the desire to target potential customers more accurately, it has also been applied to more general social phenomena. Ginsberg et al. (2008) showed that Google search term frequency patterns provided good insight into geographically-specific flu infection rates. This model was later deployed in the swine flu epidemic, to support developing countries with underdeveloped public health capabilities. Choi and Varian (2009b) and Choi and Varian (2009a) link similar data to socioeconomic outcomes like the unemployment rate.

Twitter has quickly become a major force in this trend towards more and richer data. The service began operation with the first message in 2006. By 2007, it could boast of traffic estimated at 60,000 messages per day and rising.(Douglass, 2007) As of September 2010, Twitter reported 175 million registered users generating 95 million messages, or "tweets", per day.[2] Researchers have exploited the real-time nature of Twitter content to predict movie box-office performance (Asur and Huberman, 2010), detect earthquakes (Sakaki et al., 2010), and investigate brand identification and sentiment.

For political scientists, tweets may offer an alternative or supplement to political polling

---

[2] A given "tweet" consists of only 140 characters, including spaces. Tweets originate from users, who may decide whether to broadcast to whomever chooses to look (public accounts) or only to authorized users (private accounts). Most users do not appear to bother with private tweets; indeed, much of Twitter's attraction is the ability to attract extremely large masses of "followers" via public broadcast.

Users "follow", or subscribe, to other users' tweets, similar to the functioning of an RSS syndication service for news or weblogs. The distribution of followers appears to follow a semi-log profile, with a few users counting millions of followers while most users have relatively few. It's not clear how many of the millions of tweets generated per day actually get read.

Finally, "tweets" may be either original content or re-tweets of content from others. Several studies have used re-tweet frequency as a means of estimating the relative approval for a given message. However, this remains difficult as re-tweets may function as a medium for publicizing particularly appalling or distasteful comments by important figures.

as a means of gauging voter sentiment. Because they push out information from voters, the semantic content of tweets may be more likely to reflect actual voter attitudes than polling. Polling has a well-known problem of framing, whereby respondents' answers to questions are heavily conditioned by ancillary characteristics of the polls themselves. As an information-push medium, tweets may be less influenced by these factors and as such represent a more indicative survey of voter sentiment. Pursuing this possibility, Tumasjan et al. (2010) show that mere counts of tweets provide a highly accurate polling medium for party election outcomes in Germany. In the United Kingdom, Tweetminster (2010) claim to have predicted the 2009 General Election outcomes with 90% accuracy at the national level, and with 69% accuracy at the seat level. Their work followed on similar attempts in Japan in 2009. O'Connor et al. (2010) show that Twitter feeds for the presidential candidates correlate with polling outcomes and may be a leading indicator of polling performance. Finally, Conover et al. (2011) show that semantic analysis of tweets provides insight into patterns of partisan polarization and intra- and inter-partisan communication.

3 The 2010 Congressional Election

The 2010 elections to the House of Representatives produced a historic swing in party control. The Republican party gained 63 seats in the House, giving it a large majority and producing the largest electoral swing since 1938. The 2010 elections also marked a watershed year for Twitter penetration among Congressional candidates. Political consultants Burston-Marsteller report that over 60% of Congressional members in both houses of Congress had accounts as of fall 2010.

Anecdotes suggest that Twitter played an important role in political messaging during the election. After the 2008 Republican defeat, then-Republican National Committee chairman Michael Steele suggested that part of the Republican failure stemmed from insufficient engagement with new social media. Subsequently, microblogging services like Facebook and Twitter became important communication channels for opinion-makers in the Republican party in particular. Perhaps most infamously, former Vice Presidential candidate Sarah Palin used Twitter to encourage a vigorous post-2008 response by Republicans in swing districts via her 'Don't retreat, reload' messaging. But more mainstream sources, such as Top Conservatives on Twitter (denoted by their hashtag `#tcot`) also sought to improve messaging through social media. The decentralized nature of the Tea Party was thought possible, in large part, due to this and other kinds of decentralized communication channels.

4 Research design

Predicting election outcomes from twitter message volume can be treated as a supervised learning problem. Given a set of tweets associated with a candidate or district, the election provides two forms of coding: a discrete case, or win/loss; and a continuous case, or the percentage of vote received by the candidate. The problem then becomes determining whether it is possible to generate a prediction algorithm that accurately maps patterns of political communication on Twitter to patterns of wins and losses or vote share.

We can abstract this problem into three stages: Data acquisition to filter potentially relevant messages from millions of daily tweets; data aggregation and cleaning to turn these individual messages into a format suitable to map onto a much smaller set of electoral outcomes; and algorithmic learning to train and test predictive algorithms.

4.1 Data acquisition

Twitter provides no *a priori* way of identifying politically-relevant tweets. A very sophisticated approach to this would require gathering and tagging millions of potential tweets.

Table 1: Rates of Twitter ID penetration by candidate category

	Chal.	Incumbent	D	R
No	33.94	51.8	51.11	33.19
Yes	66.06	48.2	48.89	66.81

This data set could then form the basis of a filter that could be applied to the entire twitter feed.

Lacking the resources for such an approach, this paper instead proceeded from the assumption that the most relevant tweets would contain the full name of the candidate for election. Candidates were identified from the list of Congressional election races provided by the OpenSecrets project at the Center for Responsive politics. Only the two-party Democrat/Republican ballot was identified. The final two-party candidate pair was confirmed after the primary date for a given race had passed. The final candidate list contained only races with two confirmed candidates. Special elections were omitted.[3]

Candidate names were then confirmed through research on the candidate's website. In cases where the formal name differed from the colloquial name–as in "Jack" instead of "John"–the colloquial name was used instead. Any remaining ambiguities were then checked against the record of legislative races maintained by the New York Times. Candidate twitter IDs were identified from either the candidate's website or the TweetCongress project. The resulting data set contained 916 candidates for House and Senate races. From table 4.1, it's clear that Twitter IDs were more common among challengers than incumbents, and among Republicans than Democrats.

---

[3] Several coding issues were identified after the fact. Maryland District 4 was incorrectly assigned 2 Democratic candidates rather than a Democrat and a Republican. In another case, one of the two candidates was listed as Independent. Finally, the candidate list was accurate only at the time of first data gathering. Subsequent changes to race dynamics–either early drop-out or replacement–were not reflected in the data gathering strategy.

Acquisition of the Twitter data began in late September 2010, after all but the Hawaii primary elections had occurred. Nightly queries for each candidate in the data set were submitted to the Twitter Search API.[4] Queries submitted to the API were structured as follows: the query itself consisted only of the colloquial first and last name of the candidate. Searches were performed once every twenty-four hours. The first data acquisition searched for data as far back as possible into the Twitter archive. Subsequent acquisitions searched for tweets from the prior day. In this way, as complete a history as possible of the Twitter feed related to the Congressional nominees was constructed on an iterated basis.

The initial data acquisition produced approximately 58,000 tweets for the 916 candidates in the data set. Subsequent days produced anywhere from 8,000-20,000 tweets. The behavior of message volume over time is shown in figure 1. As is apparent from the scaling, the volume of tweets per district varied dramatically, both within the two chambers of Congress and between them. Senate races tended to receive much more attention via Twitter relative to the number of races, than did House races. Much of that, however, was due to the extraordinary presence of the Delaware Senate race, where Christine O'Donnell attracted significant scrutiny and Twitter traffic volumes.

Electoral results were taken from the New York Times website, at `http://elections.nytimes.com`. Electoral outcomes were coded as 1/0 based on vote percentages and cleaned by hand in the few instances where a candidate won with a plurality rather than a majority of the vote. Those data were merged with the Twitter data on the basis of state, district, and party. The data were subsetted to include only those districts for which both candidates had some twitter volume, and in which no candidate ran unopposed.[5] That

---

[4] For full documentation, see `http://apiwiki.twitter.com/`. The search API was preferred to the streaming API due to the more straightforward query mechanisms and the ability to access the full Twitter data feed. Full access to the entire twitter stream requires special permission from Twitter and significant bandwidth resources.

[5] Even though the candidate list had been pre-screened to include only two-candidate races, subsequent changes rendered some of that invalid. In some instances, the New York Times reported only "other" as a

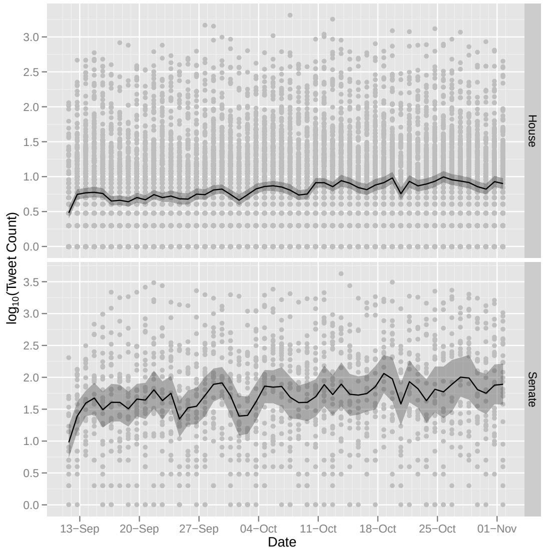


Figure 1: Tweet acquisition volume over time for House and Senate races. Shaded areas show bootstrapped 95% confidence intervals around the district mean.

resulted in 356 House races out of a total of 435.

4.2 Data aggregation and coding

For the House of Representatives race, this acquisition strategy produced a dataset of approximately 230,000 tweets for 356 races or 712 candidates.[6] Data aggregation and coding followed a three-step strategy. First, extraneous information in the text–English stopwords, retweet tags, URLs, and usernames–were removed. Second, politically relevant proper names were re-coded. The name of the candidate was replaced by a generic party-candidate tag in tweets that originated from queries for that candidate. National figures–

---

contesting candidate; in others, the major-party candidate dropped out, leaving only minor parties as challengers.

[6] Post-acquisition examination of the original total of 245,256 led to two discoveries. First, one candidate shared a name with a kicker for the New Orleans Saints football team, leading to the acquisition of a large number of tweets about the 2011 football season. Second, Philadelphia Phillies player Cliff Lee somehow made it into the dataset, for reasons unclear. Removal of tweets related to these two issues resulted in the 230,000 total.

President Barack Obama, Speaker of the House Nancy Pelosi, and Senate Majority Leader Harry Reid–were re-coded as `PresidentDummy`, `SpeakerDummy`, and `LeaderDummy`. These recodings were done for three reasons. First, they prevented proper name diversity from overwhelming subsequent text analysis. Second, recoding permitted cross-district comparisons of twitter messages. Third, recoding of national figures potentially extends the relevance of the final model to future elections where the President, Speaker, or Majority Leader might differ.[7]

This encoding strategy implies an assumption of semantic equivalence among districts. Consider, for instance, a Democratic candidate in Little Rock, Arkansas and another in San Francisco, California. Under this recoding strategy, those two candidates are both coded as `DCandDummy`. As section 4.5 will show, that will lead to treating the twitter "conversations" in each district as exchangeable. On its face, this assumption seems implausible given the diversity of race and district characteristics. But it permits the aggregation of tweets within districts and comparison across them, and so trades uniqueness for statistical power.

With the data cleaning concluded, the tweet collection was transformed into a document-term matrix of bigrams (unique 2-consecutive-word combinations), in which each tweet was represented as the count of unique bigrams it contained. This very large, sparse matrix (230,000 messages * 500,000 bigrams) was the summed into a set of district-level "documents" that counted the term frequency by district for the entire general election campaign.

Summing tweets into districts implicitly meant summing across time. How to weight messages by time thus became an open question. Four different weightings, all on the

---

[7] The one exception here occurred in cases where Nancy Pelosi appeared in tweets resulting from queries for her own name or that of her challenger. In those cases, Nancy Pelosi was coded similar to any other candidate.

interval $(0, 1]$ were tried:

1. A uniform weight that did not distinguish by time
2. A linear weight
3. A quadratic weight that loaded term frequencies close to the election much higher than those further away
4. A sigmoidal weight whose inflection point was the median date of the general election campaign

Finally, the vocabulary used for the district-level documents was filtered in a 2-stage process. First, very sparse terms were removed. Terms present in less than 1% of cases were removed for use in the win/loss predictor; and in less than 2% for the vote share predictor. Second, a set of parallel data sets was created by filtering these datasets by the mean term frequency-inverse document frequency (Tf-Idf) measure of each covariate.[8] This was done only for the win / loss predictor, at a threshold value of 0.0005. In aggregate this set of filtering steps generated document-term matrices wtih 356 districts and 1000-13,500 terms depending on the strictness of the filter.

4.3 Descriptive statistics

National and district-level tweet volumes were very unevenly distributed. Figure 4 shows the log-scaled distribution, indicating that orders of magnitude separated the highest-volume districts from the lowest. Most districts got only a few hundred tweets in total over the course of the election, while a few districts received ten to twenty times as many.

---

[8] Tf-Idf is commonly used in information retrieval to filter out both high-frequency, very common terms, and low-frequency, uncommon terms. The goal is to identify those terms which are both important and distinguishing within the corpus. Formally, for a term t^* in document $d_i, i \in 1...I$, $tfidf_{t^*,d_i} = tf(t^*, d_i) * idf(t^*)$, where $tf(t^*, d_i)$ is the term frequency for term t^* in document d_i, and $idf(t^*) = log(I/\sum_{i \in I} t^* \in d_i)$.

Table 2: House electoral outcome prediction accuracy using relative tweet volume. $N = 356$.

	By Pct. Tweet	
	Loss	Win
Elec. Outcome		
Loss	120	73
Win	68	95

However, figure 2 shows that tweet volume was only loosely correlated with the closeness of the race–very close races received only marginally more Twitter attention than less contested races.

Relative tweet volumes are only weakly predictive of outcomes. Figure 3 shows the relationship between final vote outcomes by candidate and district to the candidate's share of tweet volume for that district. The Pearson correlation is 0.39, with a p-value of 0 against a null hypothesis of no correlation. But a simple binary assignment rule to election outcome whereby a district tweet share of 50% or more is taken to be a "win" generates an accurate prediction for electoral outcomes in only 60% of the cases, as shown in table 2.

4.4 Topic modeling for content discovery

To establish the pattern of conversation in the entire corpus of messages, we turn to two different forms of topic modeling. As described initially by Blei et al. (2003), topic modeling treats documents–in this case, collections of tweets by district–as having been derived from topics that have pre-defined term distributions. Those topics, in turn, constitute latent variables that can be discovered from the empirically observed term distributions in a document collection. Assigning documents to topics on the basis of the terms they con-

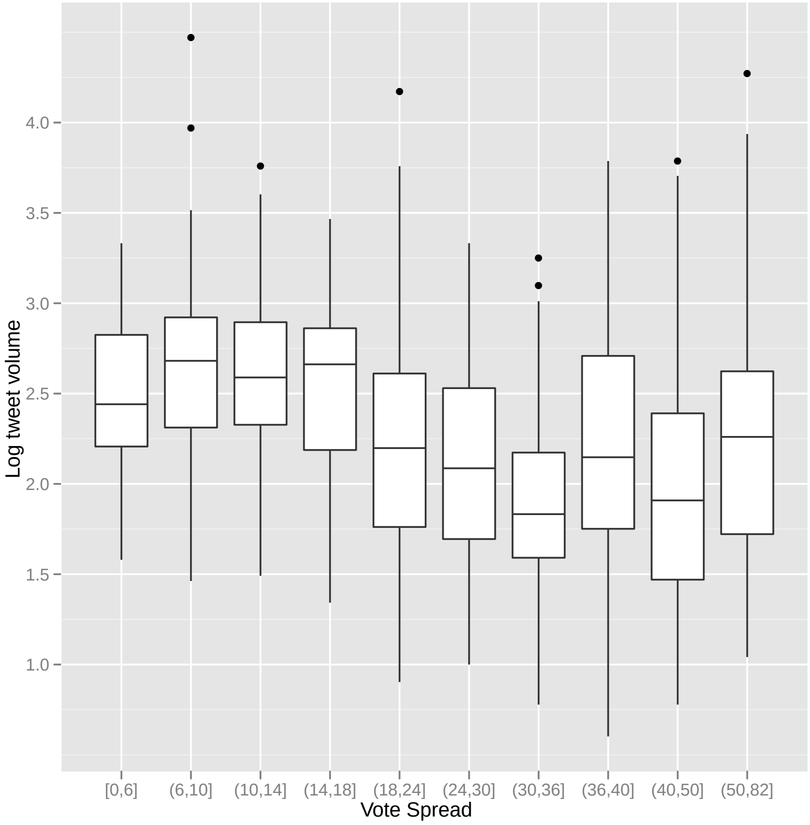


Figure 2: District-level tweet volume by degree of contestation as measured by final vote spread.

tain thus becomes a problem of Bayesian inference based on the term distributions in each document.

Figure 6 shows the distribution of a 5-topic correlated topic model (Blei and Lafferty, 2006a) run on the entire twitter corpus. For this purpose, each district's entire term distribution was treated as a "document" and assigned to one of five topics based on the term distribution. Of particular interest is the breakdown between discussions of issues–like health care or the stimulus bill–and more candidate- ore movement-focused discussions.

Dynamic topic models (Blei and Lafferty, 2006b) provide an alternative way to look

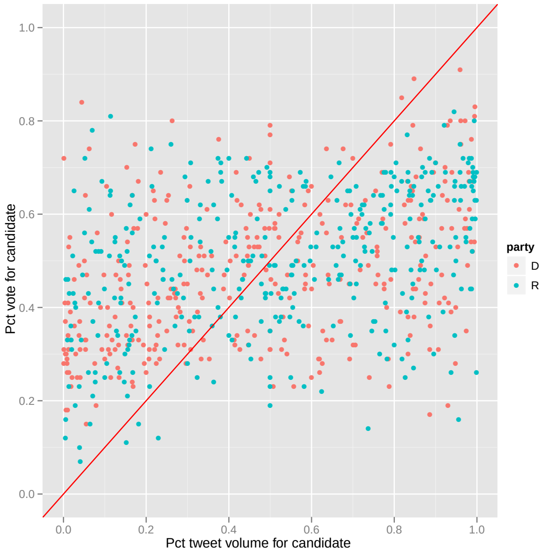


Figure 3: Candidate vote percentages versus share of district-level tweet volume for House and Senate races.

at not only the topics, but also their evolution through the course of the campaign. For example, a discussion of political issues may initially emphasize health care, but later shift to the stimulus bill. Dynamic topic models provide a means of extracting both the latent topics from the corpus of documents, and their internal variation over time.

Figures 7 and 8 provide two different views of a 5-topic dynamic model run for each district over each week in the general election. As with the static model, we can see the separation between topics emphasizing issues like the health care reforms or stimulus package, and topics focusing more on candidates or races. Of particular interest is the relatively greater presence of the "issues" topic later in the campaign, while the frequency of topics focusing on candidates fell off in importance.

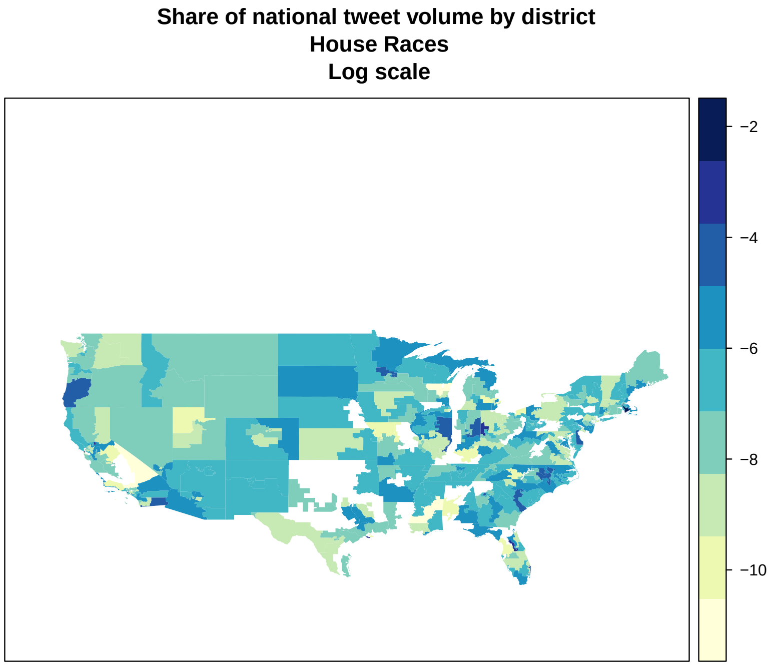


Figure 4: Share of national tweet volume by district. Untracked districts shown in white.

4.5 Machine learning and prediction

Identifying and exploiting variation in this document corpus to predict election outcomes can be treated as one of two kinds of machine learning problems. Give win/loss outcomes, the problem resembles a document classification task where the classes–"win" and "lose"–were provided by the voters. For vote share outcomes, the problem instead is a classic regression problem.

Regardless of whether the process involves regression or classification, we would like the algorithms themselves to discover the best predictors from the data. This implementation is complicated by a very large number of features–up to 80,000 unique words or

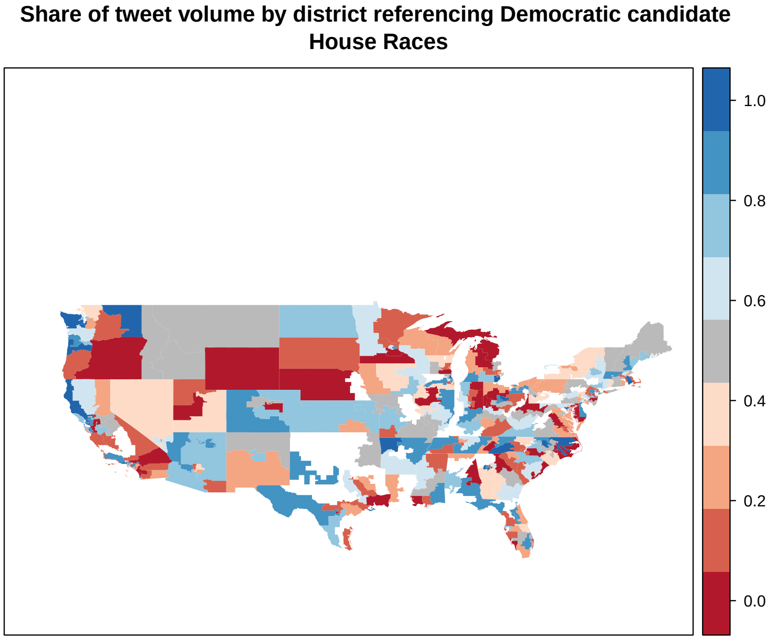

Share of tweet volume by district referencing Democratic candidate
House Races
1.0
0.8
0.6
0.4
0.2
0.0

Figure 5: Democratic candidate share of tweet volume by district. Untracked districts shown in white.

600,000 possible bigrams across all 356 "documents" in the data set. This requires the selection of appropriate algorithms to deal with very sparse, high-dimensionality matrices.

All learning processes employed the SuperLearner ensemble machine learning algorithm.(Polley and van der Laan, 2010) For either classification or regression tasks, the SuperLearner evaluates a library of specified algorithms and generates a synthetic prediction algorithm as the weighted combination of the library that minimizes a cross-validated risk score for predictive accuracy. Van Der Laan et al. (2007) show theoretically that the SuperLearner is at least as accurate as the most accurate single algorithm in the library. Empirically, they demonstrate through simulation studies that the accuracy of the ensem-

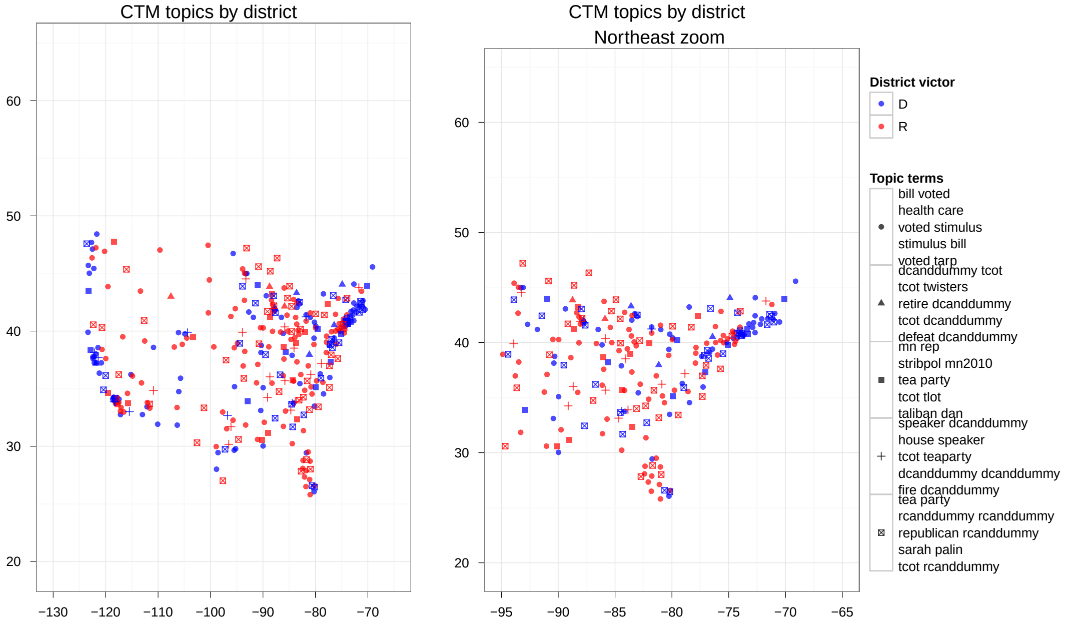


Figure 6: Geographical distribution of topics derived from a 5-topic coordinated topic model for the entire general election campaign. Points represent the geographic centroid by House district. The legend shows the top five terms for each topic.

ble algorithm can often surpass the best single library candidate algorithm.

All predictive models were built on the basis of a training data set and evaluated against an out-of-sample testing data set. The training set consisted of 285 randomly selected districts, and excluded the 71 testing districts. The SuperLearner used 10-fold (win-loss) or 15-fold (vote share) cross-validation for identification and selection of the ensemble learner on the basis of the training data set. Table 3 shows the comparative accuracy for the binary and vote share predictors in both the in-sample training data set and the out-of-sample testing dataset.

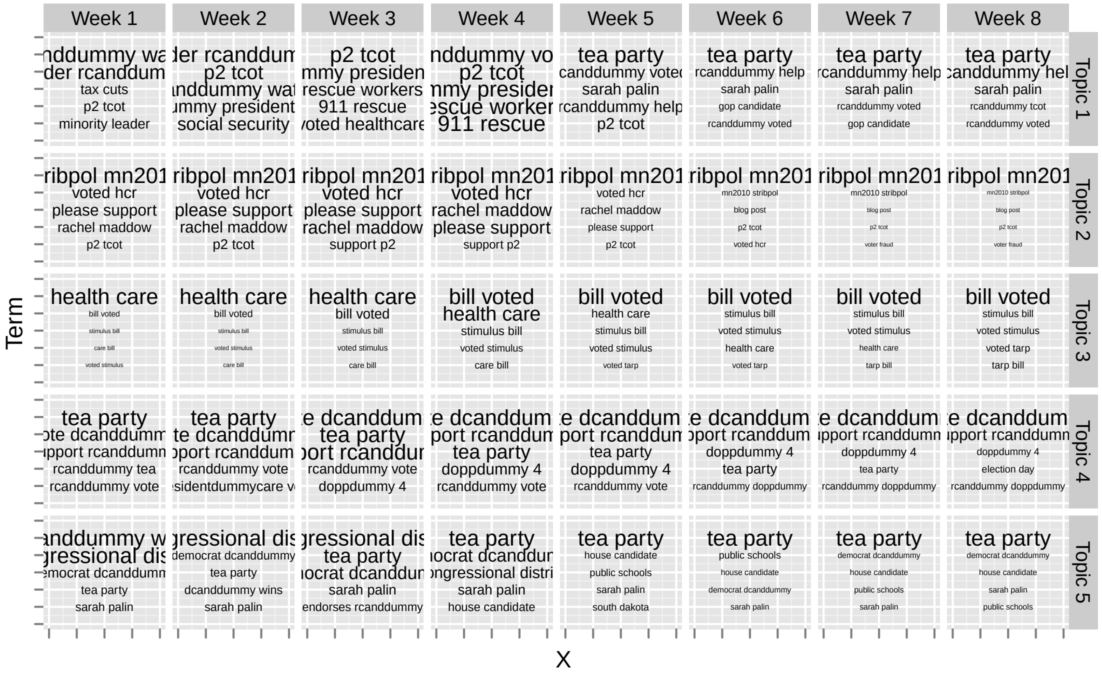


Figure 7: Top terms by week and topic for a 5-topic dynamic topic model. Each document was the aggregated term frequencies for a district for one week during the general election. Term size reflects its relative weight by period.

4.5.1 The discrete case: win/loss outcomes

For the sparse binary classification case, the SuperLearner library consisted of random forests, support vector machines with a radial basis function and either C- or nu-classification, and elastic net regression with lasso pre-screening.[9] The best estimator achieved an in-sample accuracy rate of 100%, and an out-of-sample rate of 92%, using linearly-weighted term-frequencies filtered by the mean Tf-Idf score.

Several studies (O'Connor et al., 2010; Ginsberg et al., 2008; Asur and Huberman, 2010; Choi and Varian, 2009a,b) have shown that social media can act as leading indicators.

---

[9] Screening in this and other cases refers to the use of lasso to pre-identify likely predictors within the entire dataset. This reduces the number of covariates then passed to the prediction algorithm itself.

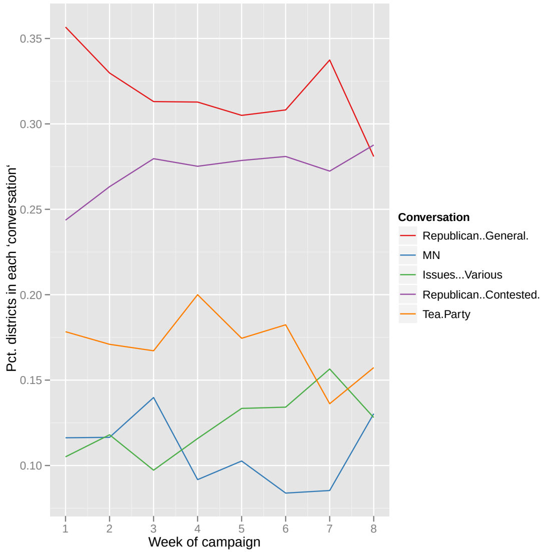


Figure 8: Count of districts assigned to topics by week in the general election campaign.

Preliminary evidence suggests that this is true for the predictors considered here as well. Accuracy rates may be stable up to three weeks prior to the election, suggesting that the terms of conversation have stabilized by then.

4.5.2 The continuous case: vote share outcomes

In contrast to the discrete prediction case, which resembled document classification, the problem of predicting vote share resembles regression. Again, however, dimensionality and algorithm selection remain problematic. A hybrid SuperLearner library was employed to take advantage of algorithms otherwise unsuited to the high-dimensionality case. Bayesian regression, gradient boosting, linear regression, polynomial splines, and step regression were all used on a dataset pre-screened for significant covariates using lasso regression. These algorithms were supplemented by sparse partial least squares, ridge regression, support vector machines, and lasso applied to both a lasso-screened

Estimator	Sample	Selection	Uniform	Linear	Quad.	Sigmoid
Binary	Testing	Tf	73	87	89	86
	Training	Tf	100	100	100	100
	Testing	Tf-Idf	–	92	87	89
	Training	Tf-Idf	–	100	100	100
Pct Vote	Testing	Tf	86	85	83	83
	Training	Tf	94	94	95	94

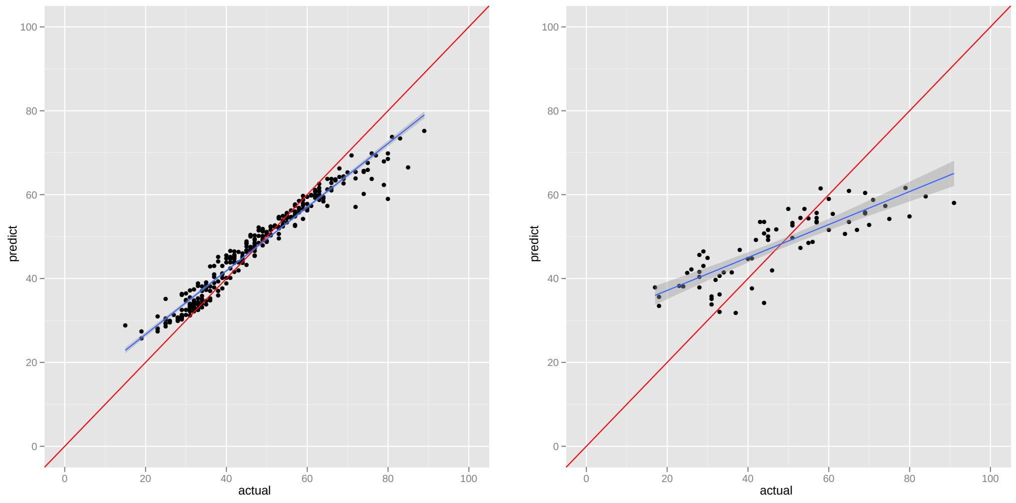
Table 3: Comparative win / loss accuracy rates for discrete and continuous predictors. Selection refers to the initial vocabulary selection mechanism: Tf used term frequency only, removing terms present in fewer than 1% (discrete) or 3% (continuous) of districts. Tf-Idf removed terms with a mean Tf-Idf value of less than 0.0005.

dataset and to the full data set. Finally, the random forest method was used on the unscreened dataset. Lasso screening of covariates usually reduced the approximately 4200 unique bigrams (at 3% sparseness) to 10-40 bigrams depending on the characteristics of the cross-validation subsample.

Table 7 shows the final composition of the ensemble. The mixture of screened and unscreened predictors suggests that the strategy of pre-screening in order to use algorithms less suited to high-dimensionality cases was rewarded.

Figure 9 shows that the resulting ensemble predictor achieved reasonably good predictive accuracy in the training dataset, and somewhat worse accuracy in the training dataset. The fitted linear regression curves clearly show that, on average, the prediction algorithm under-predicts high vote shares and over-predicts low ones.

Conversion of the predicted vote shares to binary win/loss outcomes using the 50% cutpoint as the decision rule shows that, even with misprediction, the rate of accuracy remained high. The uniformly-weighted training dataset achieved 94% accuracy in predicting win/loss outcomes; and the algorithm generated accurate predictions in 86% of the held-out testing data.



(a) Training data set, N=285 (b) Testing data set, N=71

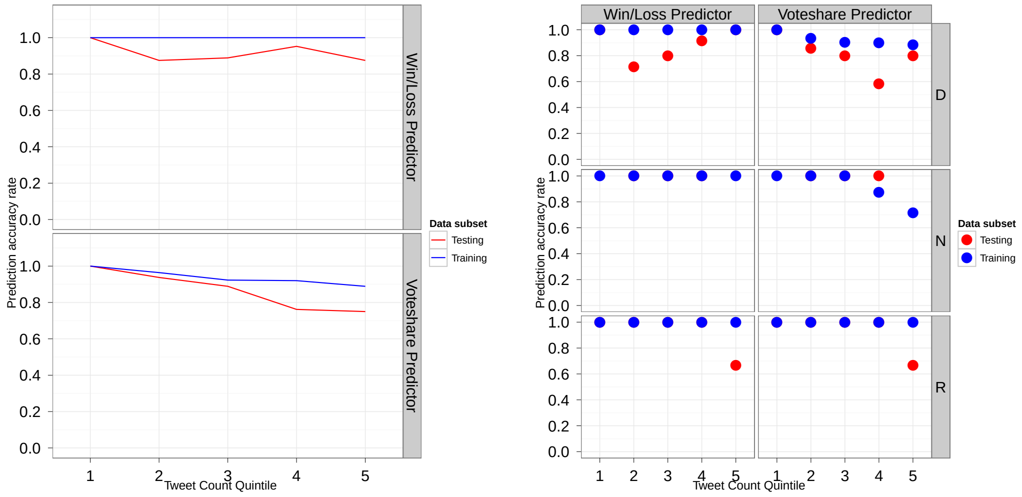
Figure 9: Predictive accuracy for vote share in training and testing datasets. Red lines indicate where 1:1 correspondence would lie. Blue lines indicate the fitted result of OLS regression of the predicted vote share on the actual vote share.

5 Conditional performance metrics

To check whether the predictors as trained on the aggregate data display disparate results on subsets of the data, prediction accuracy was checked against two factors: the incumbency status of the district, and the volume of tweets received by that district. The first is particularly important for the future accuracy of the trained predictors. Because the 2010 House election produced a strong anti-Democratic-party wave, the predictor may over-weight features predictive of Republican victories. The latter begins to address the question of how large a message volume is required before a district begins to converge on the representative behavior of all districts.

Figure 10(a) shows predictor accuracy conditional on district tweet volumes. Note that the accuracy of the test subset remains roughly consistent across the entire volume range for the binary predictor, but degrades with very large volumes for the continuous predictor. Figure 10(b) shows that result further broken down by the party of incumbency. True to the Republican bias of the election, the predictor displays much higher accuracy rates in districts where Republicans were incumbent compared with Democrats.

Finally, figure 11 illustrates that both predictors fared less well with highly contested districts as measured by the final vote spread between winner and loser. This may indicate the need to train a separate predictor on historically close races, where features prominent in safer seats may have less relevance to expressed sentiment in the twitter feed.



(a) Performance conditional on tweet volume

(b) Performance conditional on tweet volume and incumbent party.

Figure 10: Predictor performance conditional on district characteristics.

6 Performance compared to other electoral projections

Comparison against longstanding district-level electoral prediction models provides the ultimate test of the value of a twitter-based approach. Six different electoral predictions–the Cook Political Report, the Rothenburg Political Report, Congressional Quarterly, Larry

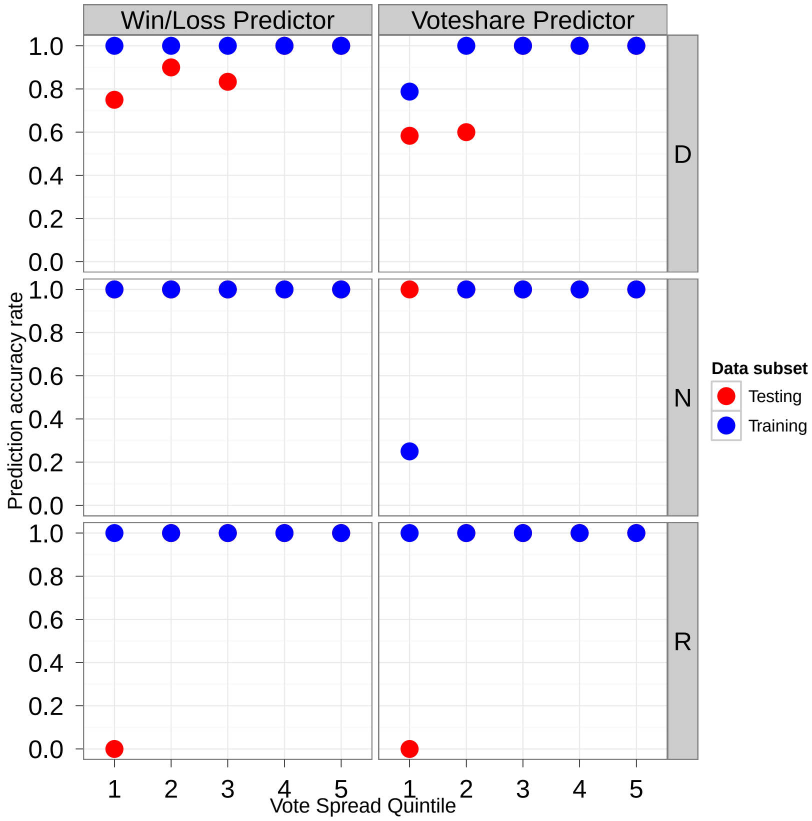


Figure 11: Predictor performance conditional on final district vote spread

Sabato's Crystal Ball, Real Clear Politics, and Nate Silver's 538.com at the *New York Times*– were selected as baseline models. All but 538.com made predictions in self-assessed "competitive" House districts. Those predictions were ranked on a scale of "Likely" / "Leans" Democrat to "Likely" / "Leans" Republican. For purposes of comparison, "Likely" and "Leans" were treated as definite predictions (i.e., "Likely Democrat" was coded as "Democrat"). "Tossup" districts were coded as NA or "no prediction".

Figure 12, reproduced in table 5, shows the accuracy rates for each forecast compared to the election outcome. Results are shown for both the testing subset of the twitter data, and the entire twitter dataset. In all cases, the comparison models did not offer predictions in many of the 356 districts contained in the full data set. The accuracy estimates therefore use only those districts for which both the baseline model and the twitter models made predictions. Notably, for the full-sample predictions, the twitter model for binary win/loss outcome prediction did better than any of the baseline models. In the out-of-sample testing data, the win / loss predictor performed as well as Congressional Quarterly, but performance otherwise degraded against the other forecasts.

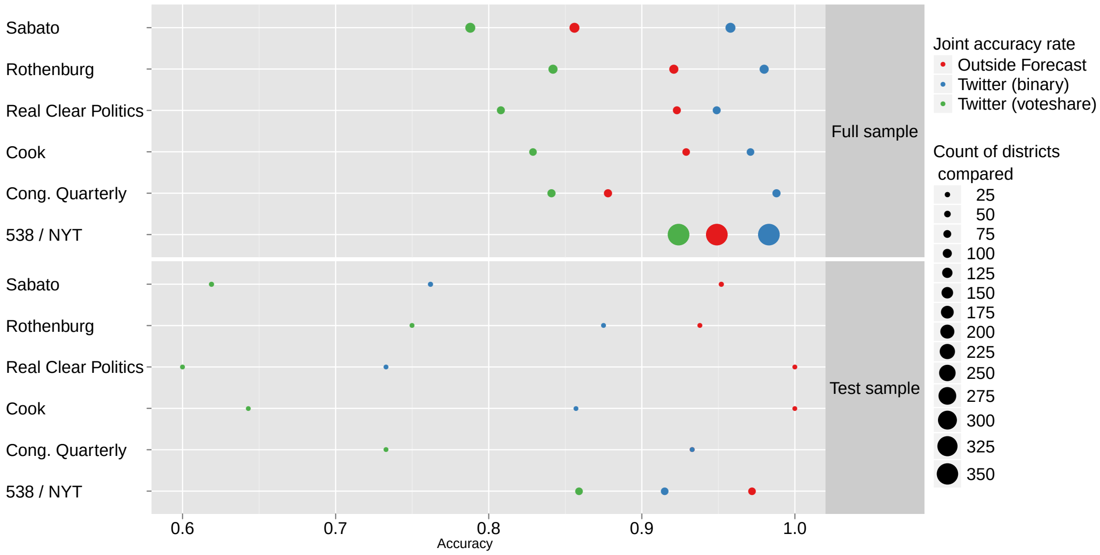


Figure 12: Predictive accuracy for baseline and Twitter-based predictors.

Finally, each of the comparison models except Sabato treat some races as "too close to call" or tossups. Table 4 shows the predictive accuracy of both the discrete and binary twitter predictor models for these tossup races. Two conclusions are immediately apparent. First, in all cases, the twitter models do better than chance. Second, the discrete model is always more accurate than the vote share model.

Table 4: Predictive accuracy for the Twitter predictors in races identified as tossups by other predictive models

	N	Twitter (discrete)	Twitter (voteshare)
Cook	48	0.938	0.729
Rothenburg	18	0.833	0.556
CQ	36	0.889	0.667
Real Clear Politics	40	0.975	0.750

7 Performance in expectation

The results presented to this point have relied on out-of-sample tests on a fixed split of training and testing data. Here we present some evidence that these results may represent the upper bound of accuracy for the estimator. In expectation, we may revert to the same win / loss predictive accuracy as the simple rubric of predicting the incumbent would win. Inspection of the most influential terms from the random forest algorithm presented in figure 14 suggests why this might be. The relative importance of terms containing both an office signifier and a party signifier appear to allow the algorithm to establish, on the basis of text alone, the party of the incumbent in the race. The algorithm may thus simply build a simple heuristic to establish incumbency from the data at hand and predict, with some modification, who will win from that. Removing these bigrams with a set of incumbent-party dummy variables, reduces the influence of these terms. Removing the data altogether significantly reduces the out-of-sample accuracy rate of the predictor.

8 Discussion

Using only the semantic content of twitter messages referencing House candidates, it was possible to predict out-of-sample election outcomes correctly in 92% of cases. That success rate improves on methods that only use relative tweet volumes. It also avoids the complication of using semantic tagging, which in any case suffers from ambiguity about the specific relationship of the semantic tag and the candidate. It also improves on an approach using supervised topic modeling, which achieves 75% accuracy in both the in-sample and out-of-sample cases. Whether these accuracy estimates will hold for future Congressional elections is unclear.

Several questions remain. First, the 2010 election was in many respects unique. The Tea Party, the depth of the recession, the newness of Twitter as a communication medium, and other factors may all render the prediction algorithm used here irrelevant for future races. It remains to be seen whether this algorithm can do better than measures like relative tweet volume in *a priori* prediction of race outcomes.

Second, this paper did not attempt to link the semantics of Twitter communication to the social networks in which it occurs. Thus we don't know if the information here is predictive because it influenced voters, or was an expression of already-existing voter sentiment. Given the dataset, it should be possible to map semantics onto the originator and follower data to determine how sentiment propagated through space and time.

Third, the data collection method here has several obvious selection biases. Presumably, anyone sufficiently motivated to sign up for Twitter and write about candidates is a reasonably high-information voter compared to the average. But that sample is further biased by differential rates of technology adoption and literacy and other factors. The relatively low marginal cost of using Twitter–the accounts are free and the service is available via a variety of electronic media–may mitigate this issues somewhat. But it also may break any supposed connection between the motivation to tweet and the motivation to actually vote. Finally, this survey has only picked up on the sentiment of those who tweet, not those who follow the tweets.

Nevertheless, these accuracy estimates may also represent a lower bound for the potential accuracy of a twitter-based predictor of electoral outcomes. The learning strategy used here ignores a significant amount of information present in twitter messages. It does not make any adjustment for retweets, which may signify more important messages. It does not recognize the presence of web URLs in the tweet; this therefore ignores both the fact that the tweet is sharing information beyond the content of the tweet itself; and the content of that information and its implication for the candidate. Finally, the learning strategy does not take into account the network of users that exist around each candidate or race. The characteristics of that network–its strength, density, frequency of interaction, or change over time–may reflect changes in the prospects for one candidate or another. Given that this information is ignored in the learning strategy above, the strategy itself may represent a minimalist lower-bound to what is possible with a twitter-based predictor.

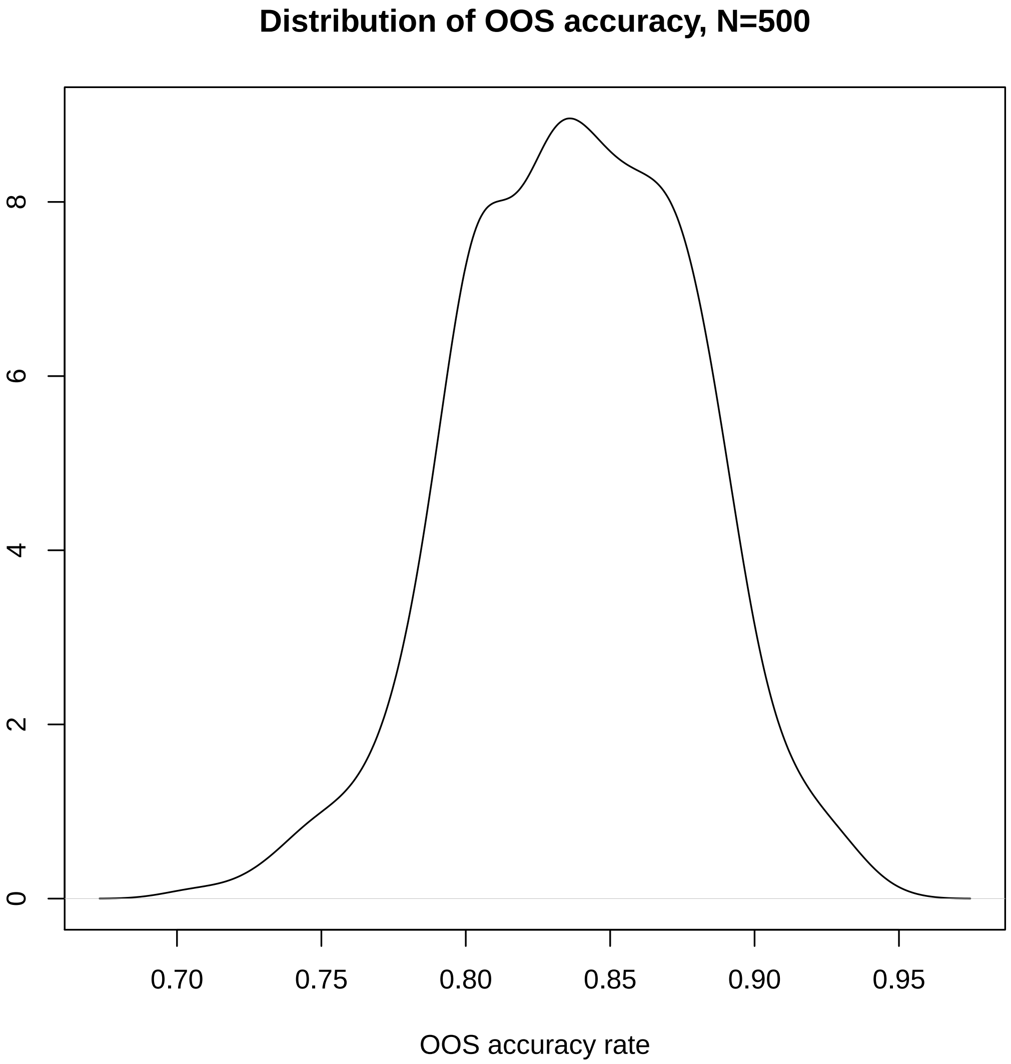


Figure 13: Bootstrapped out-of-sample predictive accuracy for the SuperLearner algorithm. Estimates based on out-of-sample prediction from 500 independently learned algorithms based on 500 independent training/testing splits.

A Appendix

Table 5: Predictive accuracy for the binary and vote share Twitter predictors, compared to actual race outcomes. Denominator in all cases is the count of districts for which both the twitter predictor and the outside forecast made predictions.

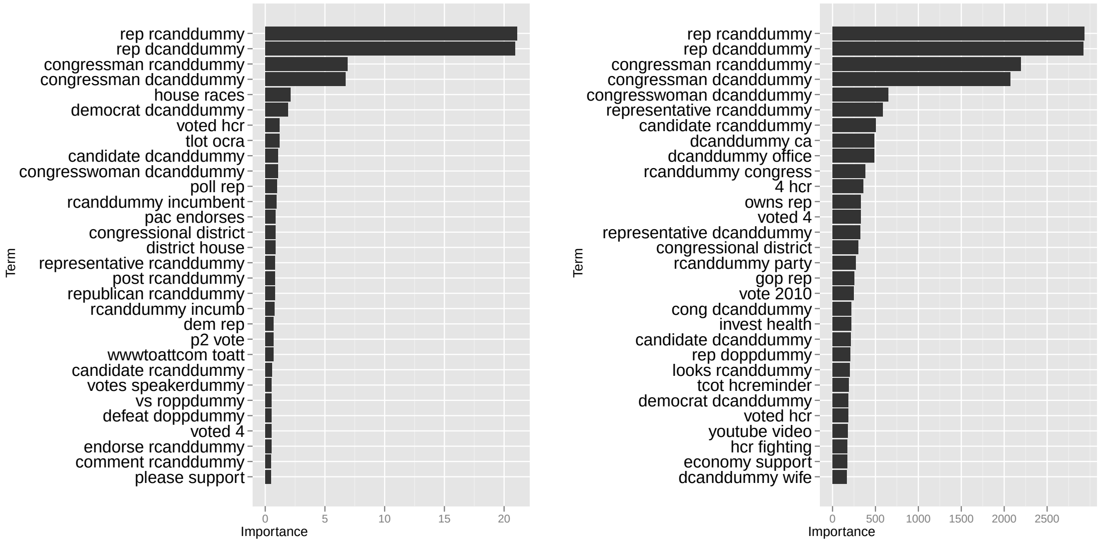
	N	Outside predictor	Twitter predictor (binary)	Twitter predictor (voteshare)
Test sample				
Cook	14	1.000	0.857	0.643
Rothenburg	16	0.938	0.875	0.750
Cong. Quarterly	15	0.933	0.933	0.733
Sabato	21	0.952	0.762	0.619
Real Clear Politics	15	1.000	0.733	0.600
538 / NYT	71	0.972	0.915	0.859
Full sample				
Cook	70	0.929	0.971	0.829
Rothenburg	101	0.921	0.980	0.842
Cong. Quarterly	82	0.878	0.988	0.841
Sabato	118	0.856	0.958	0.788
Real Clear Politics	78	0.923	0.949	0.808
538 / NYT	356	0.949	0.983	0.924

Table 6: Algorithm weightings for the discrete prediction output from SuperLearner

	Model Weighting	Risk
Screened SVM	0.033	0.271
Screened elastic net reg.	0.057	0.258
Random Forest, maxtrees=116, nodes=5	0.910	0.124

Table 7: Algorithm weightings for the vote share prediction output from SuperLearner

	Model Weighting	Risk
Screened boosting	0.120	133.945
Sparse Partial Least Sq.	0.004	1783.652
Random Forest, maxtrees=116, nodes=5	0.875	103.114



(a) Win/loss estimator

(b) Vote share estimator

Figure 14: Relative term importance for Random Forest algorithm in the binary and continuous predictor ensembles.

References

Asur, S. and Huberman, B. (2010). Predicting the future with social media. *Arxiv preprint arXiv:1003.5699*.

Blei, D. and Lafferty, J. (2006a). Correlated topic models. *Advances in neural information processing systems*, 18:147.

Blei, D. and Lafferty, J. (2006b). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Choi, H. and Varian, H. (2009a). Predicting initial claims for unemployment benefits. Working paper, Google, Inc.

Choi, H. and Varian, H. (2009b). Predicting the present with Google trends. Working paper, Google, Inc., Mountain View, CA.

Conover, M. D., Ratkiewicz, J., Francisco, M., Goncalves, B., Flammini, A., and Menczer, F. (2011). Political polarization on twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Douglass, N. (2007). Twitter blows up at sxsw conference. *Gawker.com*, March 12.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Kirkpatrick, D. (2011). Tunisia's inner workings emerge on twitter. *The New York Times*, 22 January(A6).

O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. (2010). From Tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pages 122–129.

Polley, E. C. and van der Laan, M. J. (2010). Superlearner. Working paper, Divison of Biostatistics, University of California Berkeley, Berkeley, CA.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Stone, B. and Cohen, N. (2009). Social networks spread defiance online. *The New York Times*, 15 June:A11.

Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*.

Tweetminster (2010). Is word of mouth correlated to general election results? the results are in. *tweetminister.co.uk*, 12 May 2010.

Van Der Laan, M., Polley, E., and Hubbard, A. (2007). Super learner. *Statistical applications in genetics and molecular biology*, 6(1):25.