Voting with your Tweet: forecasting elections with social media data

Mark Huberty markhuberty@berkeley.edu University of California, Berkeley

Introduction

I demonstrate that the application of ensemble machine learning techniques to the Twitter message feed can generate predictive algorithms for Congressional election outcomes that achieve greater than 85% accuracy. That accuracy rate remains stable up to two weeks prior to the election, and compares favorably with other district-level forecasts such as Congressional Quarterly.

Social Media and Social Reality

Researchers have shown that internet transaction data can support highly accurate predictions of real-world behavior:

- Ginsburg et al (2008):
  Regional influenza rates
- Choi & Varian (2010):
  Macroeconomic aggregates

As of late 2010, **Twitter reported 175 million users generating 95 million messages** per day. Twitter has been shown to reflect aggregate political outcomes:

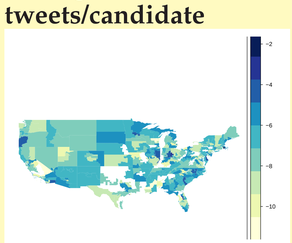
- Tumasjan et al (2010):
  National party vote share in Germany
- O'Connor et al (2010):
  Presidential candidate approval polls

Question: Can Twitter accurately predict disaggregated outcomes like district-level elections?

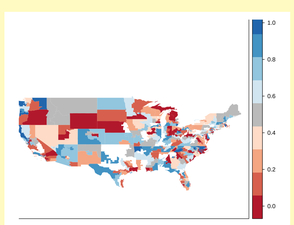
Data Collection

Daily Twitter query for all congressional candidate names during the 2010 general election season

Averaged **8,000-20,000 tweets/day** and **300 total tweets/candidate**



Log $_{10}$ share of national tweet volume by district



Democratic share of tweet volume by district

Machine Learning

Approach:
Treat this as a **supervised learning** problem mapping "documents" to district outcomes

- Apply **natural language techniques** to district-level aggregation of tweets to generate word pair features
- Use the **SuperLearner ensemble ML algorithm** (Polley et al 2010) for both feature selection and weighting

  Background: build a synthetic prediction algorithm from a library of candidate algorithms via minimization of cross-validated errors with NNLS.
- Tailor algorithm library to deal with extreme sparsity ($N \ll p$)
- Train on election outcomes for 80% of data (271 districts), evaluate on held-out 20%
- Predict either **Democratic victory** or **Democratic vote share** by district

From tweets to features

Natural Language Pre-Processing in 6 steps:

1. Remove English stopwords, URLs, usernames, retweet tags, etc.
2. Replace candidate names with party-specific placeholders (*dcanddummy, rcanddummy*)
3. Replace leadership names with placeholders (*PresDummy, SpeakerDummy, LeaderDummy*)
4. Collect tweets by candidate into district-level "documents"
5. Restrict dataset to contested districts
6. Parse into a **document-term matrix of bigrams**

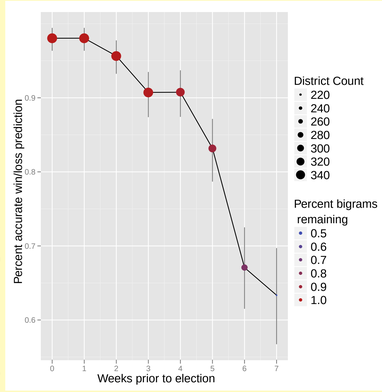
House Elections Result:
A matrix of **356 districts** containing over **240,000 tweets** with **600,000+ unique bigrams**.

Terms present in fewer than 1% of cases (binary) or 3% (voteshare) were dropped, reducing the term count to less than 17,000.

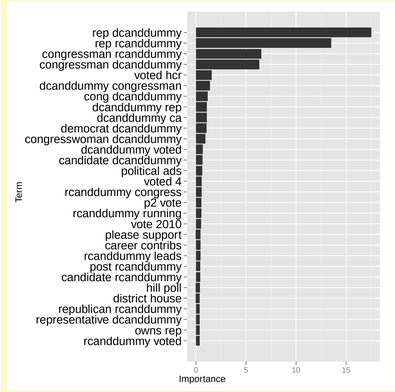
Win/loss prediction: 90% out-of-sample accuracy

Results:

- 100% in-sample accuracy
- 90% out-of-sample accuracy
- **Twitter is a leading indicator**: Full-sample predictions 95%+ accurate two weeks prior to election
- Trained algorithm reliant on **random forests** (92%) and **elastic net** regression (8%)



Accuracy of binary forecast with deletion of tweets close to election (**full sample**)

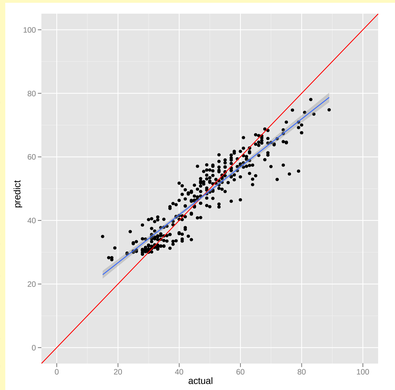


Top 30 most important terms, random forest algorithm

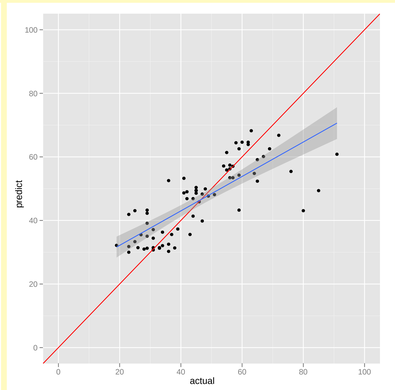
Vote share prediction: 90% out-of-sample accuracy

Results:

- 92% in-sample win/loss accuracy at 50% cutpoint
- 90% out-of-sample accuracy
- Trained algorithm reliant on **boosting** (42%), **random forests** (32%), and **MARS** (14%), **Ridge** (7.5%), and **sparse partial least squares** (4%) regression



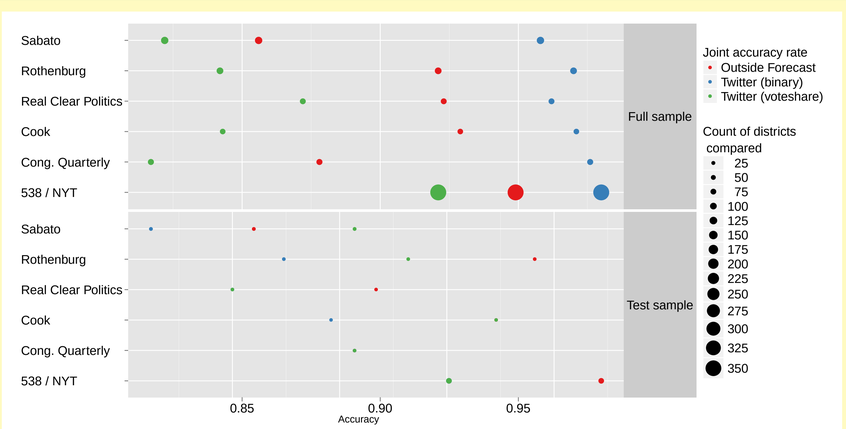
Predicted vs. actual Democratic vote share, **training** data



Predicted vs. actual Democratic vote share, **testing** data

Accuracy comparable to Congressional Quarterly

- Out-of-sample accuracy rates **equal those of CQ** and compare favorably to other forecasts
- Win/loss **forecasts for "tossup" races 83-95% accurate**
- Twitter-based forecasts **perform well by ROC measures** (recall, F, error) against all but the 538 forecasts



Twitter forecast accuracy compared to mainstream forecasts