Understanding Election Candidate Approval Ratings Using Social Media Data

Danish Contractor
IBM Research - India
Vasant Kunj
New Delhi, India
dcontrac@in.ibm.com

Tanveer A. Faruquie
IBM Research - India
Vasant Kunj
New Delhi, India
ftanveer@in.ibm.com

ABSTRACT

The last few years has seen an exponential increase in the amount of social media data generated daily. Thus, researchers have started exploring the use of social media data in building recommendation systems, prediction models, improving disaster management, discovery trending topics etc. An interesting application of social media is for the prediction of election results. The recently conducted 2012 US Presidential election was the "most tweeted" election in history and provides a rich source of social media posts. Previous work on predicting election outcomes from social media has been largely been based on sentiment about candidates, total volumes of tweets expressing electoral polarity and the like. In this paper we use a collection of tweets to predict the daily approval ratings of the two US presidential candidates and also identify topics that were causal to the approval ratings.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Applications**]: Data Mining; I.2.7 [**Natural Language Processing**]: Text Analysis

Keywords

Social media, election prediction, social network, regression, granger causality

1. INTRODUCTION

The rise of social media has led to an unprecedented opportunity to exploit social interaction and social expression. The successful application of online campaigns in both the 2008 Presidential Elections in the US and 2012 London Mayoral elections has shown the importance of online political opinions. Social media has been used in the past to try and predict election outcomes [5, 4, 1, 3]. In this paper we study how tweets in an election can be used to determine topics that are of "importance" to an election and contribute to the approval ratings of election candidates. We formulate the problem as a time series regression problem where the approval rating for each candidate is dependent on the bigrams mentioned in posts by "supporters" of the election candidates. For our work we made use of data collected on the 2012 US Presidential Elections.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. ELECTIONS OUTCOME CAUSES

2.1 Identifying "supporters"

Using two dictionaries[1] (one for each candidate), that contained keywords most associated with that candidate we classified users into support bases of either candidates based on the number of matches of tokens from their twitter profile descriptions that matched in the dictionaries. For example, a user describing himself as "right leaning" is more likely to be a republican supporter than a democrat. In case the profile descriptions were missing or there was no clear class of support, the user was termed to be a "neutral user".

Once we identified the who the most frequent tweeters in our data set supported, we built a regression based prediction model as described in the next section.

2.2 N-gram based regression model

Algorithm 1 Build prediction model using n-grams occurring in the data

Let A_d be the approval rating for a candidate on date d and m be the set of most frequent bigrams in the data
for each date d in the collection **do**
  for each of the two political candidates **do**
    Let the set of users who support candidate c (or are neutral), be U_c
    for each user $u_c \in U_c$ **do**
      for each ngram $ng \in N$ **do**
        Value of predictor feature f_c for candidate = number of occurrences of bigram ng in posts made by user u_c
      end for
    end for
  end for
end for
Build linear regression model using the set of f_c's as predictors and approval ratings A_d as predictions.

Using a regression model based on the volume of bigram mentions by supporters of the candidates, the approval ratings for each of the candidates is predicted.

2.3 Causality between topics (bigrams) and ratings

Granger causality [2] is widely used method to determine causal relationships from time series data. A time series x is said to be be granger causal for another time series y if

[1]Dictionaries were created by manual inspection of twitter user profile descriptions of most frequent tweeters.

building a regression model combining the two series gives better predictions than a model built using only the time series y.

$$y_t \approx A.\vec{y_{t-1}} + B.\vec{x_{t-1}} \quad (1)$$

$$y_t \approx A.\vec{y_{t-1}} \quad (2)$$

If on applying a t-test or an F-test on the predicted outcomes from the two regression models above, shows a significant improvement in prediction when both time series, then the causality x is said to granger cause y. Since we use multiple features in our models, performing a pair-wise granger test using each feature would be computationally expensive and therefore we use the Lasso Granger method for causality determination.

2.3.1 Lasso Granger method

The Lasso algorithm for linear regression performs variable selection using the L_1 penalty term to obtain a sparse estimate of the coefficient vectors β. The variable selection can be obtained by solving the following optimization problem:

$$\min_\beta \sum_p^P || \ y - \sum_{i=1}^N \beta_i x_{ip} \ ||^2 + \lambda \ || \ \beta \ || \quad (3)$$

where λ is the penalty parameter that determines the sparseness of β. The series x is said to cause y iff β is a non-zero vector. We made use of Lasso Granger regression implementation provided by Matlab[2] for our experiments.

3. EXPERIMENTS AND RESULTS

3.1 Data set

We used the presidential candidate opinion poll released by Gallup. Gallup conducts opinion polls concerning political, social, economic issues and regularly publishes poll outcomes. For the 2012 US Presidential elections, Gallup collected data[3] from approximately 3,050 registered voters by asking whom they would vote for if the elections were held at that time and reported seven day rolling averages for the opinion poll ratings. Using the freely available twitter4j[4] Java API library we collected over 37 million tweets between Sept 7, 2012 and Nov 7, 2012.

3.2 Training

Using the poll ratings collected from Gallup as the predicted variable, and using the bigram features as predictors we trained regression models in a seven day window. Thus, data from the last seven days was used to train a regression model and predict the poll rating for both presidential candidates on the eighth day.

3.3 Causal analysis of bigrams

Using the lasso granger regression model based on the predicted variables and the bigram features we identified the bigram features that were causal for the predictions. We found 227 bigrams that were causal for Barack Obama during the elections and 183 bigrams for Mitt Romney. Table 1 shows some examples bigrams that were found to be causal for the two candidates. From the list of causal bigrams it can be seen that bigrams related to "taxes", "job growth", "osama bin laden", " secretary clinton", "benghazi attack', "health insurance", "debt" etc' were detected as those contributing to the approval ratings.

For Barack Obama	For Mitt Romney
american_soil_	acceptance_letter_
auto_bailout_	al_qaeda_
bin_laden_	arab_spring_
business_owners_	birth_certificate_
climate_change_	bush_americans_
clinton_dnc_	created_jobs_
fiscal_cliff_	cut_deficit_
fix_economy_	dnc_obama_
foreign_policy_	democratic_national_
job_growth_	dnc_speech_
hillary_clinton_	equal_pay_
trillion_debt_	failed_policies_
obama_hillary_	fiscal_cliff_
taxes_obama_	health_insurance_
small_business_	jobs_report_
obama_care_	laden_troops_
obamas_economic_	mainstream_media_
obama_benghazi_	obama_apologizes_
martin_luther_	obama_lazy_
convention_speech_	obama_lied_
dnc_speech_	obama_unemployment_

Table 1: Examples of causal bigrams found for the two candidates

103 out of the 227 causal bigrams for Barack Obama mentioned his name while 14 causal bigrams mentioned his opponent Mitt Romney. However, 83 of the 183 causal bigrams for Mitt Romney mentioned his opponent Barack Obama, while his own name figured only 14 times in bigrams found to be causal to his approval rating. A further analysis of the mentions of Barack Obama, in the causal bigrams of Mitt Romney revealed that some of those associations were with negative sentiments, for example "obama_lied", "obama_lazy" etc were causal bigrams for Mitt Romney's. It's interesting that these bigrams were discovered without any explicit sentiment detection algorithm.

4. REFERENCES

[1] J. E. Chung and E. Mustafaraj. Can collective sentiment expressed on twitter predict political elections? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.

[2] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, Aug. 1969.

[3] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[4] E. T. K. Sang and J. Bos. Predicting the 2011 dutch senate election results with twitter. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, 2012.

[5] C. William and G. Gulati. What is a social network with facebook and vote share in the 2008 presidential primaries. In *Annual Meeting of American Political Science Association*, 2008.

[2]http://www.mathworks.in/help/stats/lasso.html
[3]http://www.gallup.com/poll/150743/Obama-Romney.aspx
[4]http://twitter4j.org/