Predicting US Primary Elections with Twitter

Lei Shi
R&D, Opera Solutions
12230 El Camino Real Suite 330,
San Diego, CA, US
lshi@operasolutions.com

Neeraj Agarwal
R&D, Opera Solutions
Express Trade Towers, Plot 15-16,
Floor 5-6, Noida (UP), INDIA
neeraj.agarwal@operasolutions.com

Ankur Agrawal
R&D, Opera Solutions
Express Trade Towers, Plot 15-16,
Floor 5-6, Noida (UP), INDIA
ankur.agrawal@operasolutions.com

Rahul Garg
R&D, Opera Solutions
Express Trade Towers, Plot 15-16,
Floor 5-6, Noida (UP), INDIA
rahul.garg@operasolutions.com

Jacob Spoelstra
R&D, Opera Solutions
12230 El Camino Real Suite 330,
San Diego, CA, US
JSpoelstra@operasolutions.com

Abstract

Using social media for political analysis is becoming a common practice, especially during election time. Many researchers and media are trying to use social media to understand the public opinion and trend. In this paper, we investigate how we could use Twitter to predict public opinion and thus predict American republican presidential election results. We analyzed millions of tweets from September 2011 leading up to the republican primary elections. First we examine the previous methods regarding predicting election results with social media and then we integrate our understanding of social media and propose a prediction model to predict the public opinions towards Republican Presidential Elections. Our results highlight the feasibility of using social media to predict public opinions and thus replace traditional polling.

1 Introduction

Microblogging provider Twitter, Facebook, and Google+ have become very popular communication tools for Internet and Mobile users. These platforms enable more than friendly chatter and individual expression; they facilitate remarkably diverse and broad participation while accelerating the formation of effective collaborations. Authors of those messages write about their life, share opinions on a variety of topics and discuss current issues.

Because of a free format of messages and an easy accessibility of microblogging platforms. Internet users tend to shift from traditional communication tools (such as traditional blogs or mailing lists) to microblogging services.

As more and more users post about products and services they use, or express their political and religious views, microblogging websites become valuable sources of people's opinions and sentiments. Such data can be efficiently used for marketing or social studies [1]. In [2], the authors tried to predict the real-world outcomes with Twitter and reported that the volume can be used to predict several kinds of consumer metrics such as forecasting box-office revenues for movies before their release. On another hand, instead of predicting the future, Choi et al. [3] tried to predict the present with google trends and reported that the good trends can be used to predict retail sales in " Motor vehicle and parts dealers". In addition, researchers have reported that the publicly available social media data can be used to predict flu epidemics [4], stock market trends [5], housing market trends [3] and politics [6].

However there are conflicting claims regarding predicting election results by using Twitter. O'Connor et al. [6] and Tumasjan et al. [7] have reported that Twitter data has a strong correlation with presidential elections. However, Gayo-Avello et al. [8] claimed that Twitter data seems to be a poor electoral predictor because of demographics bias. Since the data that were used in all of those work were collected before 2010 and there have been a dramatic change for social media since then , it is worth reevaluating the predictive ability of social media in political issues such as elections. In order to find people's opinions about political issues, we investigated the relationship between public opinion towards republican presidential candidates during the republican primary election period of 2012. Millions of tweets were collected and analyzed from Twitter website through public available Twitter API since September 2011. Twitter allows users to post short messages (up to 140 characters) that are publicly visible through the Internet. By capturing tweets mentioning each presidential candidate and analyzing the sentiments behind those tweets, we could track people's opinions about each candidate and thus predict the final primary election results.

The rest of this paper is organized as follows: First, we will briefly introduce our data collection method and evaluate the traditional methods. Second, we will discuss the geo information identification method of the tweet we used in our model. Third, we will introduce the features that we believe are correlated with the election and propose our prediction model. Third, our prediction results will be presented with comparison to traditional poll results and election results. Finally, we will make some conclusions from our experiments and propose our future research.

2 Data Collection and Description

First, we decided to do some experiments by using only the Twitter volume with and without sentimental analysis to test the claims made in [6, 7]. Twitter data were collected from September, 2011 to February, 2012.

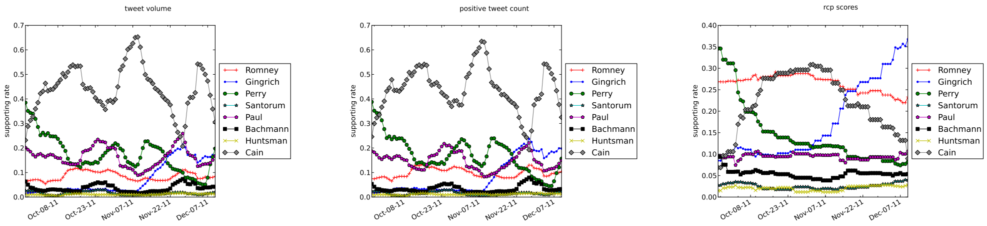
2.1 Data Sets

Twitter provides Streaming API for capturing its public data in real-time. Twitter allows 1% of all the public tweets to be sampled (Spritzer access level) for free. It sends text streams of the tweet with many other details in JSON format. In average, we received 20-25 tweets per second (TPS) from Twitter which spikes at some interval and follows a daily similar pattern. In the period between September' 2011 to February' 2012, we collected around 300 million raw sampled tweets from Twitter.

Twitter also provides an API endpoint to get relevant data based on keywords to be extracted and is also restricted to the same 1% limitation. In our experiment, we pulled out tweets for the Presidential candidates based on their full names. Over the period, totally we collected around 10 million tweets.

2.2 Storage

We used a NoSQL based database solution MongoDB to store the tweets for easy storage of JSON objects (BSON) directly into our database. MongoDB was configured in a shared environment for maximum efficiency with disk IO, network IO rates. Before storing them, several other fields were added to the JSON (ex. location, sentiment, etc.) for using them later.



(a) Prediction result with tweet volume

(b) Prediction result with positive tweets

(c) RCP scores

Figure 1: Tweet volume vs. Positive tweets vs. RCP scores

3 Initial Examinations and Observations

3.1 Methods of Prediction

In this section, predictions were computed based on Twitter volume in [7] and sentimental analysis as in [6] with slight modifications. The prediction results were then directly compared against the poll results in Realclearpolitics website [9]. Realclearpolitics website calculates poll results (RCP scores) by taking the average of poll results from several popular media and independent survey institutes. We believe the RCP scores are the closest one that we can find about public opinions towards republican presidential candidates. Thus in this paper, we are using the RCP score as our baseline and compare our prediction results with it.

Table 1: Correlation between the tweet volume and the RCP score

	Romney	Santorum	Paul	Gingrich	Bachmann	Cain	huntsman	Perry
Correlation	0.67	0.01	0.14	-0.11	0.91	0.81	-0.31	0.95

3.2 Results of the Prediction Methods

Using only twitter volume to predict the election results

Figure 1(a) shows the prediction result with only using the number of tweets mentioning each candidate and Figure 1(c) shows the RCP scores in Realclearpolitics website. The time frame is set between Sept. 17th, 2011 and Dec. 10, 2011, since Hermain Cain suspended his campaign after Dec. 10, 2011. In order to have a big picture of the correlation between the tweet volume and the public opinions, here we are using all eight presidential candidates. As we can see in Table 1, the tweet volume has a strong correlation with poll result for some candidates, e.g. Herman Cain, Rick Perry and Michele Bachmann. But there is less correlation for the other candidates. This result conflicts with what Oconnor and Tumasjan found [6, 7]. One reason is that they only evaluated two candidates in their experiments which may happen to be a coincidence. The other reason may come from twitter itself, when Oconnor and Tumasjan did their experiments, twitter had not been widely used as a platform for discussing presidential elections.

To further analyze the capability of tweets, we also evaluated the prediction results by using sentimentally analyzed tweets similar to the method used in [6, 7, 8]. As indicated in Figure 1(b), by only using sentimental analyzed tweets or the volume of positive tweets, the predicted results are very similar to using only the tweet volume. From our experiments, we had similar conclusions with Gayo-Avello in [8] that by using the volume of tweets or positive tweets, we can't predict the election result correctly. In order to predict public opinions towards republican presidential elections precisely, we need to come up with some new models. In the following sections, we will illustrate the model we built to solve this problem.

4 Geo Information Identification

Identification of the true location of the twitter users is important in order to be able to capture the public sentiments in a particular area. Especially when we come to predict election results for each state, we have to be able to identify which state the tweet comes from. However, only less than 1.5% of the tweets contain the actual location or geo information of the user. Consequently, we need to develop a mechanism to identify the location of a significant number of twitter users.

Each twitter user profile contains a 'Location' field where user can enter their location information in a free form text. However, identifying user's location from the free form text present in the 'Location' field is a challenging task. Users can provide all different sorts of information in this free form text field. The information may be some valid geographic information or the information may be invalid. The user may enter information at locality, city, state or even country level. Some people mention their favorite actor's name in the Location field. Sometimes people also use insulting language in the field to express their sentiment and sometimes they even mention irrelevant thing like 'In my bedroom' or 'Here, there, everywhere'.

```
input  : User Specified Location String $Uloc$, GMT offset present in the tweet $GMT_{in}$
output: Predicted location $< City, State, Country >$ or 'No Match'
Remove stopwords from $Uloc$;
if $Uloc$ is blank then
    return 'No Info'
end
Search the gazetteer for locations with all terms in $Uloc$ matching exactly;
if  no matching location is found then
    search the gazetteer for locations with some terms in $Uloc$;
end
Select candidates with GMT offset within +/- 1 hr of input GMT offset $GMT_{in}$;
if no candidate is selected in the previous step or $GMT_{in} == `None'$ then
    select all candidates;
end
Reject candidates where city/state name matches only partially with $Uloc$;
if no candidate is selected then
    return 'No Match';
end
Select the candidate with the highest population;
Calculate the confidence score for the selected candidate according to the scoring model;
Return the selected candidate location along with its confidence score;
```

Algorithm 1: Location Recovery algorithm

We have designed a system in Algorithm 1 that analyzes the text written in the 'Location' field by the users and identifies his/her location by searching in a gazetteer of all world cities. However, just predicting a best match location doesn't complete our solution to the problem. We also need to be precise for every prediction that we make. Any inaccuracy in our data can lead to huge biases while predicting the election results. To complete our solution, we have also designed a scoring model in Algorithm 2 that outputs a confidence score indicating how confidant the system is about the correctness of the predicted city or Location.

5 Features and Model

Choosing the right features from Twitter data is the key issue to predict election result accurately. So far, we only have seen researchers used Twitter volume with or without sentimental analysis. We believe that there must be some other features that are correlated with the election result. Below we will briefly discuss each of the feature types we used in our prediction model.

```
input  : User Specified Location String $Uloc$, GMT offset present in the tweet $GMT_{in}$,
         Predicted location ¡$City, State, Country$¿
output: The confidence score ¡$Conf$¿ in the range 0 to 1
Initial Confidence = 1.0;
if predicted city name is present in $Uloc$ while state name/code is not present then
    $Conf = Conf * 0.8$
end
if city name is present and there are more than 1 matching candidates then
    $pRatio = \frac{Population-of-selected-city}{Sum-of-population-of-candidates}$ $Conf = Conf * pRatio$
else
    $Conf = \frac{Conf}{No.-of-unique-state-candidates}$
end
if $GMT_{in} == `None'$ or GMT offset could not be retrieved for the selected candidate then
    $Conf = Conf * 0.9$
else
    if $GMT_{in}! =$ GMT offset of the Predicted Location then
        if $GMT_{in} - 3600$ ¡= GMT offset of predicted location ¡= $GMT_{in} + 3600$ then
            $Conf = Conf * 0.95$
        else
            $Conf = Conf * 0.85$
        end
    end
end
```

Algorithm 2: Scoring Model

5.1 Feature Description

1. **Tweet Volume:** the number of tweets mentioning one candidate. As described above, we found out this feature has high correlation with the public opinions for some candidates.
2. **retweet volume:** the number of retweets mentioning one candidate. Retweet is another important term in Twitter, which is replying a tweet or forwarding a tweet to one's followers. This feature is important because it indicates how many tweets mentioning each candidate are broadcasted within the Twitter network.
3. **Twitter User Count:** the number of Twitter user accounts whose tweets mentioning one candidate. We believe not only the tweet volume is correlated with public opinion, the number of twitter users who are talking about each candidate is also an important factor.
4. **Unique Twitter User Count:** the number of Twitter user accounts mentioning only one candidate. It appears to us that one twitter account may post tweets mentioning several candidates at the same time. By eliminating the tweets and twitter accounts mentioning several candidates, we could find out the exact number of twitter users who talked about each candidate.
5. **Non-promotion twitter user account I:** the number of Twitter user accounts that are non-promotion twitter accounts of the candidate. There are several presidential candidates who hired companies to create robot accounts to follow those candidates and post tweets supporting them. Those tweets and twitter accounts don't represent the general public opinions and need to be eliminated.
6. **Non-promotion tweet volume I:** the number of tweets not coming from those promotion twitter accounts.
7. **Non-promotion twitter user account II:** the number of Twitter user accounts that are not from promotional twitter accounts of the candidate and the republican party. It is similar to the above one, but for some candidates there are some Twitter user accounts which are created by the republican party to promote the candidate.
8. **Non-promotion tweet volume II:** the number of tweets that are not from promotional twitter accounts of the candidate and the republican party.

Table 2: Prediction results vs. RCP scores for three states

Candidates	New Hampshire		South Carolina		Florida	
	Predicted	RCP	Predicted	RCP	Predicted	RCP
Mitt Romney	50	48	33	28	50	46
Rick Santorum	9	12	12	17	6	13
Newt Gingrich	19	12	42	41	35	32
Ron Paul	22	28	14	13	10	7

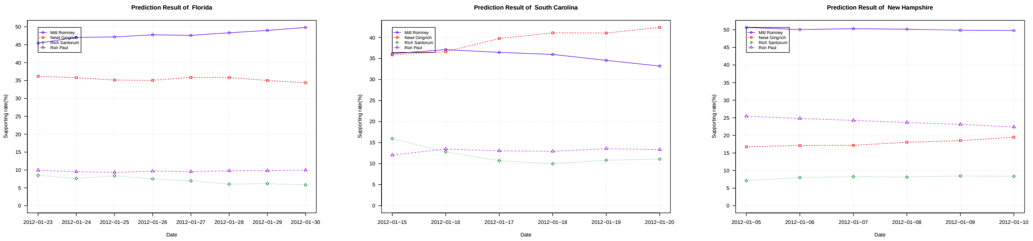


Figure 2: Prediction results vs. RCP scores for three different states for each candidate

9. **Number of tweets containing links:** the number of tweets that contain links. We found that most tweets with links are forwarding news about one candidate. So this feature is a good indicator of news coverage in the twitter network.

   In addition to the above features, we also take into account of some ratios, like the number of tweets posed per user and non-promotion tweets posted per user and eventually we have 19 features.

5.2 Proposed Model

In this paper, we develop a linear regression model to investigate the prediction problem. To control variable number and improve model specificity, we use a lasso (Least Absolute Shrinkage and Selection Operator) regression algorithm proposed by Tibshirani [10]. It imposes a constraint on the coefficients, by which the coefficients of non-relevant variables will be set to zero and thus is often used as a variable selection tool as in Tibshirani [11] and Osborne, Presnell and Turlach (2000). The central idea of the model is consistent with the nature of different features we investigate coordinately regulating a target value which in our case is the RCP score. The loss function of Lasso regression is defined as:

$$L = \sum_i (y_i - \sum_p \beta_p x_{ip]})^2 + \lambda \sum_p ||\beta_p||, \tag{1}$$

where x_{ip} denotes the pth feature in the ith datum(observation), y_i denotes the target value in this datum, and β_p denotes the regression coefficient of the pth feature. λ is a complexity tuning parameter that controls the amount of shrinkage. The larger the value of λ is, the greater the amount of shrinkage. Hence, λ should be adaptively chosen to provide an estimate of expected prediction error. The norm-1 regularizer $\sum_p \beta_p$ in Lasso regression typically leads to a sparse solution in the feature space, which means that the regression coefficients for most irrelevant or redundant features are shrunk to zero. Theoretical analysis in [12] indicates that Lasso regression is particularly effective when there are many irrelevant features and only a few training examples, which really fits our case since we only have around 100 training examples.

6 Experimental Results

6.1 State Wise Prediction

Due to the nature of American Republican Presidential election system, we have to predict the election result for each state. With our geo identification module, we could capture the tweets

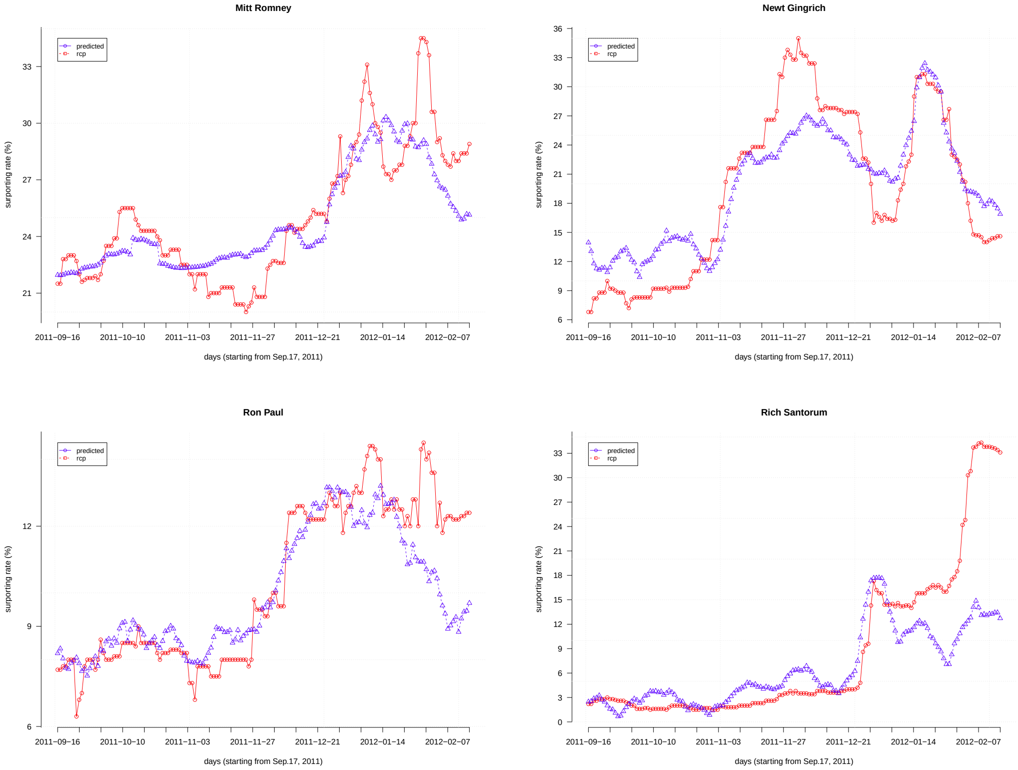


Figure 3: Prediction results vs. RCP scores nationally for each candidate

coming from each state and thus generate the corresponding features mentioned above. The target values are RCP scores for each state. Since we don't have RCP scores for each state, we can only predict the election result for three states where RCP scores are available. For each state, we collected data from Sep.17 up to the day before the primary election date of each state. The first 90% observations are treated as the training date set and the rest 10 % observations are used as the testing date set. The prediction results for the day before the election date will be used as our final prediction results for that state.

Figure 2 and Table 2 show our prediction results for those three states. As you can see, for some states like New Hampshire and Florida, the predicted results do not change too much over the week before the election day. However there is a big change for the election result for South Carolina, which reflects people's opinions have changed dramatically over the week before the election, possibly it is caused by the tax return issue of Mitt Romney during that period of time. This kind of information is really helpful to presidential candidates, since they can analyze the problems and fight back in the next coming state election.

6.2 National Wise Prediction Result

In addition, we also track people's opinions nationally over the last couple of months. Similar to the state wise prediction, we collect tweets from all states and generate the corresponding features. Again we are using 90% of the data as training set and the rest 10% as testing set. For different candidate, due to the variance of supporters, we have to build individual model for each different candidate. In Figure 3, we can see for some candidate like Mitt Romney and Newt Gingrich, the pattern of our prediction model is very close to RCP scores. But for the other two candidates, our prediction model couldn't capture the trend. The reason maybe the patterns of the training set for these two candidates are kind of flat, so it is difficult for the lasso regression model to pick up the correct features to represent public opinions.

7 Conclusions

In this paper, we have revealed that it is feasible to predict American Presidential Elections by using social media data (tweets in our case). While the results do not come without caution, it is encouraging that expensive and time-intensive polling can be supplemented or supplanted with the analysis of social media data. We argue that merely using the volume of tweets with or without sentimental analysis is not enough to capture public opinions. We need to come up with some sophisticated algorithm and model to make the prediction successfully. In this paper, we proposed our prediction model but it is by no means a prefect one. As a matter of fact, we have a long way to go to be able to capture the public trend in the real world through social media. To be able to predict election results accurately, we believe on one hand we need to understand the impact of different laxicons by using machine learning techniques. On the other hand, we need to integrate our understanding of the dynamics of political conversation in social media.

References

[1] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, may 2010.

[2] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, 2010.

[3] Hal R. Varian and Hyunyoung Choi. Predicting the present with google trends. *SSRN eLibrary*, 2009.

[4] J. Ginsberg, MH Mohebbi, RS Patel, L Brammer, MS Smolinski, and L. Brilliant. Predicting the future with social media. pages 492–499, 2010.

[5] J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. In *Journal of Computational Science*, volume 2, pages 1–8, 2011.

[6] B. OConnor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of 4th ICWSM, AAA Press*, pages 122–129, 2010.

[7] A. Tumasjan, T. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of 4th ICWSM, AAA Press*, pages 178–185, 2010.

[8] Daniel Gayo-Avello, Panagiotis T. Metaxas, and Eni Mustafaraj. Limits of electoral predictions using twitter. In *Proceedings of 5th ICWSM, AAA Press*, pages 178–185, 2010.

[9] realclearpolitics. www.realclearpolitics.com.

[10] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[11] Robert Tibshirani. The lasso method for variable selection in the cox model. In *Statistics in Medicine*, pages 385–395, 1997.

[12] Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *In ICML*, 2004.