From Total Hits to Unique Visitors Model for Election's Forecasting

Diego Saez-Trumper
Universitat Pompeu Fabra
Barcelona, Spain
diego.saez@upf.es

Wagner Meira Jr.
U. Federal de Minas Gerais
Belo Horizonte, Brasil
meira@dcc.ufmg.br

Virgilio Almeida
U. Federal de Minas Gerais
Belo Horizonte, Brasil
virgilio@dcc.ufmg.br

1. INTRODUCTION

Using Internet to predict elections has been a topic of interest for different fields. Researchers from Google have showed an approach employing user's queries on that search engine [4]. Other site, The Daily Beast, has create an "Election Oracle" [1], scanning 40,000 blogs and social media sites and applying a sentiment analysis to made their predictions. In both these case, the predictions are expressed as a likelihood of winning and not the total amount candidate votes or percent expected. This makes sense because they have applied their methodology to U.S.A elections which are based in a two-party system where one candidate won and the other lose. An multi-party approach was proposed by Tumasjan et al [3], they state that is possible to predict the result by counting the number of Twitter mentions of Political Parties and Candidates. They have tested this idea in 2009 German elections obtaining a similar accuracy of traditional election polls. However, all these methods require an important span of time to be implemented. Moreover, recent studies claim that these kind of techniques could not replace the traditional pools Limits of Electoral Predictions using Social Media Data,[2] setting, among other things, that these algorithms do not offer a methodology to sample data. Its important to note that we are not trying to replace traditional pools, but we propose a model that could - easily - give a reasonable election forecast, based in a the big amount of information available in the Social Networks. More over, we face problem of how to sample information, proposing a algorithm based on events. In this paper we take the challenge to develop an algorithm that *(i)* can predict a election result as the percent of votes that the candidates will obtain, that means a prediction useful not only to two-party, but to a multi-party system, *(ii)* that does not require heavy crawling and *(iii)* can be applied in a shorter span of time. We have tested our algorithm to the Brazilian President Elections, 2010, showing that with small but significant modifications we reach an accurate result with only one day of Twitter information.

We tackle these challenges using the Tumasjan Model as base line, considering the Twitter mentions to candidates, but we present two important improvements: First, we use a unique visitors approach versus the total hits used in the legacy methods, meaning that we consider only one mention per user, avoiding the distortion generated by heavy users, activists or Spammers. This approach allows to improve significantly the model accuracy. Second crossing our data with elections information we identify the peaks in Twitter traffic match with important events in the election camping like TV Debates. We have showed that those events allows to make accurate predictions using data obtained in a couple of hours. Our results suggests that these events are a good source of information, not only because the amount of data (the same level of data obtained in a wider span of time does not give the same of information) but because the diversity and independence of opinions that are expressed in these moments.

2. BRAZILIAN ELECTIONS AND DATASET DESCRIPTION

The Brazilian president elections 2010 occurs in two rounds: The first round was in October 3, and the three major candidates were: Dilma Rousseff , Jose Serra , and Marina Silva obtaining 46.91%, 32.61%, 19.33% respectively. As no candidate achieved absolute majority, a second round was required. On October 31, Dilma Rousseff defeated Serra with 56.05% of votes against 43.95%. Our Dataset contains all tweets mentioning main candidates by they popular names: Dilma, Serra and Marina. We have download all the tweets with these names from July 7 to November 1 of 2010. The results were 8,249,610 Tweets from 1,041,772 different users.

3. COUNTING MENTIONS - BASE LINE ALGORITHM

A first approach to establish a base line was to apply the same method proposed by Tumasjan et al, that was count all candidate mentions. As we can see in Figure 2, the activity vary significantly at a daily base, making necessary to decide which period we will take in account. Following the idea applied in German Elections we consider 5 weeks before the election, except the last week. In our case, for the first round of the Brazilian Elections that means counting from August 14 to September 25, 2010. As measure of prediction quality we have used the Mean Absolute Error (MAE).The result was an error of 9,48 MAE. Considering that the last week before the election has a significant of traffic, we tried with

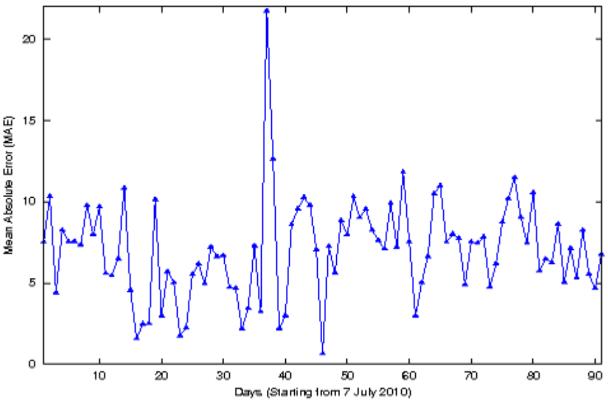

Figure 1: MAE predicting day by day.

Candidate	Mentions	%	Result	Error
Dilma	1,210,001	58.13	47.45	10.68
Serra	542,097	26.04	32.99	6.95
Marina	329,249	15.81	19.55	3.74
Total	2,081,347	-	-	-
MAE	-	-	-	7.21

Table 1: Legacy Model: Results Counting Mentions from September 1 to October 1.

a period selection that consider all the tweets done during September. Table 1 shows that the prediction improves but is still weak.

3.1 From Total Hits to Unique Visitors

As we mention before, each user may post an unlimited number of Tweets. Based on the idea of Unique Visitors used to evaluate the traffic of a Website, we propose to count only one mention per user. Applying this rule, we can compare results from legacy method (Table 1) with the new algorithm (Table 2). We can see that the prediction improves from 7.21 of legacy model to 4.07 with our technique.

Candidate	Single Mentions	%	Result	Error
Dilma	323,542	50.64	47.45	3.19
Serra	171,686	26.87	32.99	6,12
Marina	143,563	22.47	19.55	2.92
MAE	-	-	-	4.07

Table 2: Unique Visitors approach.

3.2 Event Detection

However, the solutions proposed have the problem that requires one month of sampling. We have tried to use a shortest span of time. Figure 1 shows the MAE variation on a daily prediction. We can see that the curve is not smooth, showing important daily variations. This instability suggest that could be risky make a prediction based only on one day. However, we note that exists some single days with a very good predictions (see Table 3). These good predictions match with peaks in tweets traffic. We have check these dates, and they match with important events like TV Debates and elections days (see Figure 2). These improvements could be associated with the amount of Tweets of these days, suggesting that it is only a matter of the sample size, but if we compare to the previous results based on month sample (see Table 2), we can see that even with a significant biggest sample the performance is similar. We conjecture that in event days, the diversity of users are wider providing a better picture about the reality.

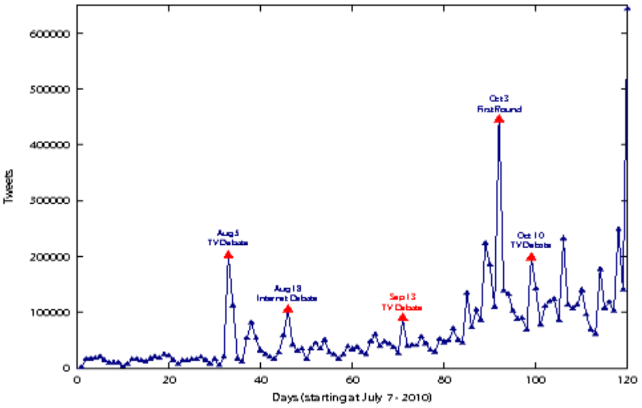

Figure 2: Peaks in tweeter traffic match with important dates (events) on the election.

Date	Event	Dilma	Serra	Marina	MAE
Aug 5	Tv Debate	48.36	34.94	16.78	1.87
Aug 18	Internet Debate	46.47	33.57	19.95	0,65
Sep 13	Tv Debate	58.42	26.70	14.87	7.31
Oct 3	1st Round Election	49,56	28,00	22,43	3.33
Oct 10	Tv Debate	57,97	43,03	–	2.84

Table 3: Predicting results with events days

4. CONCLUSIONS

We have introduced the concept of Unique Visitors given reaching a most accurate prediction, and we also found that important days in the election campaign are an important source of information allowing to obtain good predictions in a short span of time.

5. REFERENCES

[1] T. D. Beast. The election oracle. http://www.thedailybeast.com/election-oracle, 2010.

[2] D. G.-A. Panagiotis Metaxas, Eni Mustafaraj. Limits of electoral predictions using social media data. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (to appear)*. AAAI, 2011.

[3] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape. *Social Science Computer Review*, 2010.

[4] E. Wood. Searching your way to the ballot box. http://googleblog.blogspot.com/2010/10/searching-your-way-to-ballot-box.html, 2010.