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
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

Try to predict political elections using internet it is an interesting topic of research. For example, researchers from Google have showed that the queries on that search engine [4] are correlated with the elections results. Other example is the website *The Daily Beast*, they have created an "Election Oracle" [1] scanning around 40,000 blogs and social media sites and applying a sentiment analysis technique to make their predictions. In these both cases, the predictions are expressed as a likelihood of winning and not in the total amount votes or percent per candidate, this is because they have applied their methodology to U.S.A elections that are based in a two-party system, where one candidate won and the other lose.A multi-party approach was presented by Tumasjan *et al* [3], they claim that is possible to predict the result of an election counting the number of Twitter mentions of political parties and/or leaders. They have tested this idea in 2009 German elections obtaining a similar accuracy of the traditional elections polls. However, all these methods requires an important span of time to be implemented. And they also has been criticized be other studies that claims that these kind of techniques could not replace the traditional opinion-pools, [2] setting, among other things, that these algorithms does not offer a methodology to sample data.

In this paper we are not claiming that our methodology is available to replace traditional pools, but we propose a model that we consider that can give a reasonable election forecast, based in the information available on Twitter. We also study different methodologies to sample the data. Summarizing, in this paper we take the challenge to develop an algorithm that:

- (*i*) Presents the prediction as the percent of votes that the candidates will obtain, that means a prediction useful not only for two-party but for a multi-party system,
- (*ii*) does not require heavy data crawling and
- (*iii*) that can be applied in a shorter span of time.

To test algorithm, we applied it in a Twitter dataset about the Brazilian president elections in 2010. We showed that with small but significant modifications to the Tumasjan's algorithm, we are available reach an accurate result with only few days of Twitter information.

We tackle these challenges using the Tumasjan Model as base line. As Turmasjan Model, we also consider the Twitter mentions of candidates, but we present two important improvements: first, we use an unique visitors approach versus the total hits used in the legacy algorithm, meaning that we count only one mention per Twitter user, avoiding the noise generated by heavy users such as activists or Spammers. This small modification allows to improve significantly the model accuracy. Second, comparing the Twitter data with elections information we have identified that the peaks in Twitter traffic matches with important events in the election camping, for example the debates on television. We show that this 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

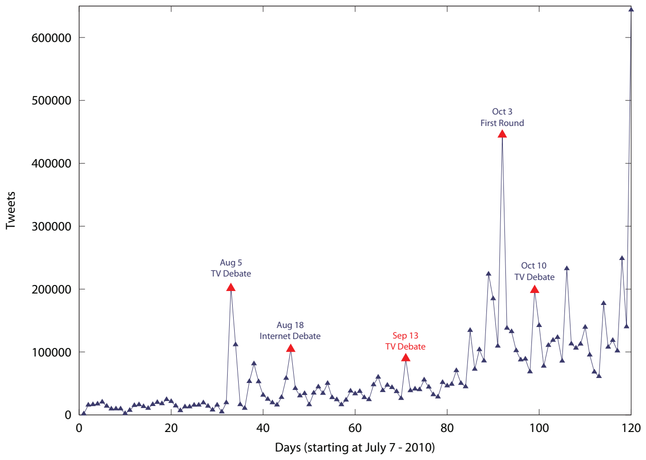

Figure 1: Peaks in Twitter traffic matches with important dates (events) on the election.

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.

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 in 2010 were in two rounds: the first round was in October 3, and the three main candidates were: Dilma Rousseff , Jose Serra , and Marina Silva obtaining 46.91%, 32.61%, 19.33% respectively. As no candidate received 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 the 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, using Twitter's streaming API[1]. The result were 8,249,610 Tweets from 1,041,772 different users.

3. COUNTING MENTIONS: THE BASE LINE ALGORITHM

A first approach to establish a base line was apply the same method proposed by Tumasjan *et al*, that was count all candidate mentions. As we can see in Figure 1, the activity have big variations day by day, making necessary to decide which period we will take in account. Following the idea applied by Tumasjan we considered the data from five weeks before the election, not including 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. This MAE is significantly bigger than the error reported in the German elections(1.65 MAE). Therefore, considering that the last week before the election have a lot of traffic, we tried with a date selection that consider all the tweets posted in September. Table 1 shows that that the prediction improved but was still weak.

3.1 From Total Hits to Unique Visitors

As we mention before, each user can post an unlimited amount of Tweets. Based in the idea of Unique Visitors used to evaluate the traffic of a Website, we proposed to count only one mention per user. Applying this rule. The result is that the Unique Visitors approach(Table 2) outperforms significantly the legacy method (Table 1), 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

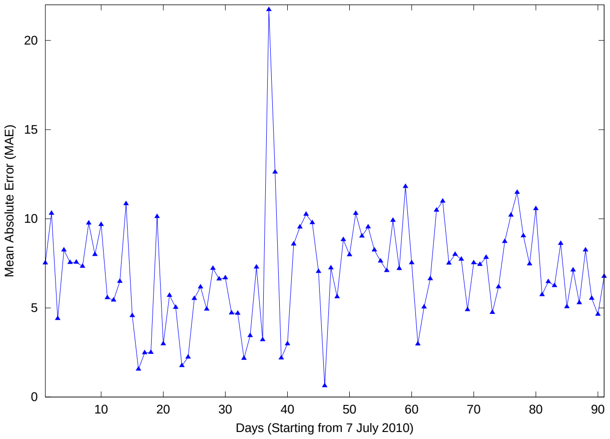

Figure 2: MAE predicting day by day.

However, the solution proposed has the problem of require one month of sampling. Hence, it is interesting try to make the prediction using shortest span of time. Figure 2 shows the MAE variation on prediction day by day. We can see that the curve is not smooth, showing important changes day by day. This instability suggests that could be risky make a prediction based only in one day. However, we note that there are some single days with a very good predictions (see Table 3). These good predictions matches with peaks in Twitter traffic. This dates matches with important events in the camping such as debates on television or elections days (see Figure 1). These improvements could be associated with the amount of Tweets on these days, suggesting that is only a matter of the sample size, but if we compare with the previous results based on month sample 2, we can see that even with a significant biggest sample the performance is similar. We conjecture that in the events days, the diversity of users are wider providing a better picture about the reality.

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	1 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 for election forecasting improving the legacy methods. We also

[1] https://dev.twitter.com/docs/streaming-api

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.