# Leveraging Candidate Popularity On Twitter To Predict Election Outcome


Manish Gaurav
Raytheon BBN Technologies
10 Moulton Street, Cambridge
MA, USA
mgaurav@bbn.com

Anoop Kumar
Raytheon BBN Technologies
10 Moulton Street, Cambridge
MA, USA
akumar@bbn.com

Amit Srivastava
Raytheon BBN Technologies
10 Moulton Street, Cambridge
MA, USA
asrivast@bbn.com

Scott Miller
Raytheon BBN Technologies
10 Moulton Street, Cambridge
MA, USA
smiller@bbn.com



## ABSTRACT

In recent years, Twitter has become one of the most important modes for social networking and disseminating content on a variety of topics. It has developed into a popular medium for political discourse and social organization during elections. There has been growing body of literature demonstrating the ability to predict the outcome of elections from Twitter data.

This works aims to test the predictive power of Twitter in inferring the winning candidate and vote percentages of the candidates in an election. Our prediction is based on the number of times the name of a candidate is mentioned in tweets prior to elections. We develop new methods to augment the counts by counting not only the presence of candidate's official names but also their aliases and commonly appearing names. In addition, we devised a technique to include relevant and filter irrelevant tweets based on predefined set of keywords. Our approach is successful in predicting the winner of all three presidential elections held in Latin America during the months of February through April, 2013.



## General Terms

Twitter, Prediction

## Keywords

Twitter, Election Prediction, Latin America, Venezuela, Ecuador, Paraguay


## 1. INTRODUCTION

Since the first tweet was sent in 2006 [1], Twitter has become one of the mostly commonly used social medium to broadcast information, thoughts and opinions. As of December 2012, Twitter boasts more than 500 million users generating over 300 million tweets everyday on variety of topics [4]. Researchers in several domains have been intrigued by the possible prediction and forecasting potential from the content generated on Twitter. There has been work demonstrating the use of Twitter in prediction of stock market [5], movie box office performance [6], pandemics [7].

We are interested in the possibility of predicting the outcome of elections from Twitter data. There have been several publications indicating the correlation between the trends and sentiment on twitter with the outcome of the election. One of the most cited work [8] indicates that Twitter is used as a platform for political deliberations and the number of tweets reflects political preferences and comes close to election polls. [9] shows that that the Twitter sentiment correlates with the polls and can capture large scale trends. Several studies have been published on Twitter's ability to predict the voting results with reasonable accuracy in the 2009 German elections [10], in the 2010 US Congressional elections [11], Dutch Elections [12], Singapore elections [14], German elections [15]. A study done in the context of Portuguese elections suggests that the volume of tweets from both news organizations and ordinary citizens followed the results from national opinions polls [16].

On the other hand, there has been staunch criticism on the studies that have claimed to predict the outcome of elections. [13] reviews the approaches and their shortcomings in the published work in the area. In particular, [13] highlights that the publications in the area show how trends on Twitter could have predicted the outcome of election after the fact rather than ahead of the election. The research also ignores several factors, such as, gender, social information, incumbency that could influence the election. [17] shows that electoral predictions using the published research methods on Twitter data are not better than chance. In addition, they emphasize the role of factors that could potentially influence elections, such as, nature of political conversation in social media, the relation between political conversation and electoral outcomes, and the way in which different ideological groups and activists engage and influence online social networks. It also reported mean absolute error (MAE) for


Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.


*SNA-KDD* '13 Chicago, Illinois, USA


Copyright 2013 ACM 978-1-4503-2330-7 ...$15.00.

the vote share of the candidates as high as 17% for Twitter volume, suggesting that without solid explanatory models such predictions are unlikely to be useful in the long run.

In this study, we examine whether Twitter data can be used to predict the presidential elections held in Venezuela, Paraguay and Ecuador in early 2013. We compute the popularity of the candidates on Twitter prior to the election and estimate the chance of their winning the elections. We developed new methods to evaluate the popularity by expanding the search to the aliases and commonly used names of the candidates. In addition, we filter the tweets that are relevant to elections based on a set of keywords.
We aim to answer the following questions:

- Is the share of Twitter messages mentioning candidate name predictive of their respective share of the vote in the election?

- Can using aliases in conjunction with candidate's full name improve the prediction of the share of their vote?

- Can filtering tweets based on specific keywords improve the vote prediction?

The rest of the paper is organized as follows. Section 2 describes twitter corpus description. Section 3 describes our methodology. Section 3.1 describes our technical approach. Section 3.2 describes leveraging alias in twitter search. Section 3.3 describes using appropriate keywords while searching aliases. Section 3.4 deals with the issue of multiple tweets sent by a single user. Section 3.5 describes our evaluation. Subsequently we present the results of our techniques. Finally, we conclude and outline future work.

## 2. CORPUS DESCRIPTION

Twitter is an online social networking service and microblogging service that enables its users to send and read text-based messages of up to 140 characters, known as 'tweets ', averaging 11 words per tweet.Twitter's public API provides only 1% or less of its entire traffic (the 'firehose'), without control over the sampling procedure, which is likely insufficient for accurate analysis of public sentiment. Instead, we collect all relevant tweets in real-time from the entire Twitter traffic via Gnip decahose [2], a commercial Twitter data provider. We collect 10% of all global tweets for any specific period of time. We approximately collected 13 billion tweets posted over September 2012 to April 2013. This comprises a roughly uniform sample of public messages, about 50 million messages per day. We analyzed a week of data prior to three Latin American presidential elections namely venezuela, paraguay and ecuador.

Tweets are represented as JSON Objects in out Twitter corpus. JSON (JavaScript Object Notation) [3] is a lightweight data-interchange format. It is a text-based open standard designed for human-readable data interchange. It is derived from the JavaScript scripting language for representing simple data structures and associative arrays, called objects. We use a JSON parser provided by java to parse the JSON object and extract text of tweet and several other attributes.

| Country | Corpus Collection Time Range | Election Date | tweets |
|---|---|---|---|
| Venezuela | April 7 - 13, 2013 | 14 April 2013 | 400m |
| Paraguay | April 14 - 20, 2013 | 21 April 2013 | 397m |
| Ecuador | Feb 10 - 16, 2013 | 17 Feb 2013 | 395m |

**Table 1: Corpus Description**

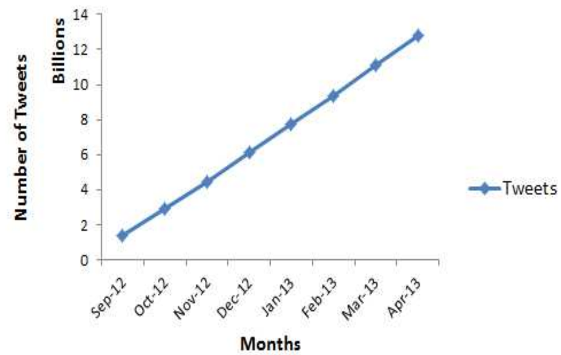


**Figure 1: Overview of our collection of tweets in the time period of September 2012 to April 2013. The data set contains approximately 13 billion tweets. The number of tweets ( approximately 50 million per day) increase almost every month. The collection is estimated to contain about 10% of the total volume of all global tweets.**

Figure 1 gives an overview of our Twitter Corpus collection. Table 1 shows amount of tweets that we used for predicting elections. Election date denotes the date at which election took place. Corpus Collection Time Range shows the time period used to analyze the tweets. We analyzed 400 million tweets for each presidential elections.

## 3. METHODOLOGY

### 3.1 Technical Approach

We employ a count based technique in which we count tweets containing specific keywords related to a particular election. We use this approach to estimate popularity for all candidates participating in the election on a per day basis. We then compute probability of candidate to win in two different ways:

#### *3.1.1 Moving Average Aggregate Probability (MAPP)*

The first approach includes computing probability of a candidate winning per day and then use the mean of probabilities in a week. We compute probability of candidate $i$ on a day $t$ as:

$$P_{i,t} = \frac{C_{i,t}}{\sum C_{i,t}} \quad (1)$$

Where $P_{i,t}$ denotes probability of candidate $i$ on day $t$. $C_{i,t}$ represents tweets mentioning candidate $i$ on day $t$. $\sum C_{i,t}$

denotes sum of all tweets mentioning all candidates participating in the election.

Finally, we compute Moving Average Aggregate Probability (MAPP) using estimated probability of a candidate $i$ on a day $t$ as a moving average over a window of the past k days:

$$MAPP_{i,t} = \frac{1}{k}(P_{i,t-k} + P_{i,t-k+1} + ... + P_{i,t-1}) \quad (2)$$

Where $P_{i,t-k}$ shows probability of candidate $i$ on a day $t-k$. We set the value of parameter $k$ as 7 to use 1 week of prior twitter data.

#### 3.1.2 Moving Average Aggregate Probability Using Counts (MAPC)

The second approach includes summing up tweet counts for a candidate for a week and then convert them into probabilites. We compute aggregate count for candidate $i$ on day $t$ as:

$$Caggr_{i,t} = (C_{i,t-k} + C_{i,t-k+1} + ... + C_{i,t-1}) \quad (3)$$

$Caggr_{i,t}$ denotes total count for candidate $i$, $C_{i,t}$ denotes count of a candidate $i$ on day $d$.

We then compute Moving Average Aggregate Probability $MAPC_{i,t}$ using aggregate counts of a candidate as:

$$MAPC_{i,t} = \frac{Caggr_{i,t}}{\sum_i Caggr_{i,t}} \quad (4)$$

### 3.2 Leveraging Candidate's Aliases

In order to explore candidate mentions in the twitter data, we search for exact candidate names in our Twitter corpus. In addition, we also search for candidate names starting with hashtags (#). We sum all mentions of a candidate in the Twitter corpus for every day until one week prior to the election. We convert popularity to probabilities by normalizing them with counts of all candidates.

We observed that a candidate can be referred to with different names (aliases) in Twitter instead of candidate's full name. For example, in the case of 2013 Venezuelan elections, Vice-president Nicolas Maduro was frequently referred to as Maduro, whereas another prominent candidate Henrique Capriles Radonski was frequently referred to as Capriles or Henrique Capriles in Twitter. We therefore missed a lot of tweets relevant to the presidential election due to the absence of candidate's full name in tweets. Table 2 shows aliases for Venezuela and Paraguay elections. Counting alias mentions of a candidate could be potentially erroneous, since presence of an alias in a tweet could be completely unrelated to the election.

In order to assess the impact of aliases, we ran two different set of experiments as described below.

We searched for two variants of each candidate's name: the full name and all possible aliases of each candidate, allowing for minor punctuation and capitalization variation. For nearly all candidates, aliases were used more often than full names. We computed a list of possible aliases for each candidate participating in the election.

| *Country* | *Candidate* | *Aliases* |
|---|---|---|
| Venezuela | Nicolas Maduro | Maduro |
| | Henrique Capriles Radonski | Henrique<br>Capriles, Capriles |
| Paraguay | Horacio Cartes | Cartes |
| | Anibal Carrillo | Carrillo |
| | Efrain Alegre | Alegre |
| | Lilian Soto | Soto |
| | Mario Ferreiro Sanchez | Ferreiro |

**Table 2: Showing different aliases used for Venezuela and Paraguay 2013 presidential elections**

### 3.3 Incorporating Keywords

A major challenge encountered by our approach is when there are multiple people with same name or alias as that of a candidate. We observed that using aliases as a candidate name could result in a lot of irrelevant tweets being used to estimate popularity for the candidate. For example, in the case of Venezuela elections, we used name 'maduro', as an alias for candidate 'Nicolas Maduro'. While that allowed us to extract a lot of tweets talking about Nicolas Maduro, we also counted a lot of tweets which didn't have any context with the presidential candidate, since Maduro is also a common name in Venezuela.

In order to alleviate this problem, we modified our approach so as to only consider tweets that contains the alias name as well as specific keywords like 'decciøn'(spanish version of keyword election) or 'election'. This allowed us to count only those tweets which express an opinion about Venezuelan election while mentioning a particular candidate. Furthermore, to broaden the search criteria, we add country name (in this case: Venezuela) as a keyword too, while looking for aliases.

Table 3 shows three sample Spanish Tweets along with their English translation for Venezuelan election. Column 'Keyword', shows the keywords present in the tweet, while column 'Candidate', shows the candidate name mentioned in twitter. The first tweet contains alias 'Maduro', as well as keyword 'elecciones'. The second tweet contains alias 'Capriles', and keyword 'Venezuela', while third tweet contains alias 'Maduro'without any keywords. As it can be see through English translations of the tweets, first and second tweets are relevant to the Venezuela elections, while the third tweet is not relevant to election. Using our selective keyword approach, we were able to select tweets one and two while disregarding the third irrelevant tweet. Its important to note that all three tweets were counted as relevant when we used the candidate name or alias matching criteria without considering any keywords.

Table 4 shows different keywords that we used for all three elections.

### 3.4 Multiple Tweets From Single User

| *Tweet* | *Keyword* | *Candidate* |
|---|---|---|
| Original Tweet: ahora **maduro** es pro-gay. bueno, que se de un besito con @dcabellor y le creemos. lo que son las **elecciones**.<br>English translation: Mature now is pro-gay. well, that a kiss with @ dcabellor and believe. what are the choices. | elecciones | Nicolas Maduro |
| Original Tweet: escuchar a **capriles** hablar me da esperanzas de que todava **venezuela** puede cambiar.<br>English translation: capriles hear talk gives me hope that STILL? can change to Venezuela. | venezuela | Henrique Capriles |
| Original Tweet: Se humilde para admitir tus errores, inteligente para aprender de ellos y **maduro** para corregirlos.<br>English translation: Be humble to admit your mistakes, learn from them intelligent and mature to correct. | none | Nicolas Maduro |

**Table 3: Showing sample Spanish tweets with English translation (using Google Translate) for Venezuela 2013 presidential elections. It also shows aliases and keywords present in the tweet.**

| *Country* | *Keyword* |
|---|---|
| Venezuela | election, elecciøn, Venezuela |
| Paraguay | election, elecciøn, Paraguay |
| Ecuador | election, elecciøn, Ecuador |

**Table 4: Showing keywords used for all three presidential elections.**

We observed that a single user can post multiple tweets on a given day which can add a reasonable bias to a particular candidate's popularity. This phenomenon is commonly known as Twitter bomb. It is generally described as the process of flooding the micro-blogging site Twitter with similar hashtags, keywords and links using same or multiple accounts. An example of a Twitter bomb analyzed in [18] described how nine fake user accounts produced 929 tweets within 138 minutes, all with a URL link to a political website, presenting negative views on the US politician Martha Coakley.

Table 5 shows statistics for Twitter users tweeting about different candidates. # tweet indicates total number of tweets for a particular candidate, while # tweet user shows total number of twitter users tweeting about the candidate. Column *recurrence* value is computed as:

$$R_c = \frac{\#tweet}{\#twitterusers} \quad (5)$$

A higher $R_c$ value shows that more Twitter users are tweeting multiple tweets about the candidate. As shown in table 5, the average number of Tweets, users tweeted per day ranged from 1.14 to 1.26 tweets in the last week prior to election for a single candidate, which is pretty reasonable. We included a single tweet per day from every Twitter user in our analysis to examine the results compared to using all tweets from a user.

In the case of Venezuela 2013 elections, 25443 users tweeted 29116 tweets mentioning Henrique Capriles Rodanski with his full name or using one of his alias ( Henrique Capriles or Capriles) as well as a specific keyword, culminating in an average of 1.14 tweets per user compared to 29213 tweets by 25592 users for Nicolas Maduro which indicates a very close election. The highest discrepancy in average tweet per user was found for Paraguayan election between Mario Ferreiro (1.26) and Horacio Cartes (1.08). The number of tweets and number of users tweeting indicates the overall popularity of the election. We found that Venezuela presidential election generated approximately 58 thounsand tweets compared to approximately 1.7k for Paraguayan election and approximately 2.3k for Ecuador elections.

| *Country* | *Candidate* | # Tweet | # Twitter Users | Recurrence |
|---|---|---|---|---|
| Venezuela | Henrique Capriles Rodanski | 29116 | 25443 | 1.14 |
| | Nicolas Maduro | 29213 | 25592 | 1.14 |
| Paraguay | Efrain Alegre | 634 | 552 | 1.14 |
| | Horacio Cartes | 707 | 651 | 1.08 |
| | Mario Ferreiro | 233 | 185 | 1.26 |
| Ecuador | Guillermo Lasso | 472 | 393 | 1.20 |
| | Rafael Correa | 1805 | 1260 | 1.20 |

**Table 5: Showing Tweet statistics for all major candidates in all three elections.**

## 3.5 Evaluation

We evaluate our results by computing the RMSE (Root Mean Square Error) between our estimated probabilities and actual outcome. It is computed as

$$RMSE_E = \sqrt{\frac{\sum_c (P_c - P_{\tilde{c}})^2}{N}} \quad (6)$$

Where $P_c$ represents estimated probability and $P_{\tilde{c}}$ represents actual outcome in probability for candidate $c$.

We determine the correlation coefficient between actual results and estimated probabilities described by the equation below:

$$Correl(X, Y) = \frac{\sum (x - \widetilde{x})(y - \widetilde{y})}{\sqrt{\sum (x - \widetilde{x})^2 (y - \widetilde{y})^2}} \quad (7)$$

Where X represents actual poll probabilities, Y is the estimated probability , and x is the mean of actual probabilities and y is the mean of the estimated probabilities. Higher correlation coefficient indicates better accuracy of our estimated probabilities

# 4. RESULTS AND DISCUSSION

We compute MAPP candidate probabilites for three presidential elections in Venezuela, Paraguay and Ecuador as defined in Equation 1 and 2 and MAPC candidate probabilities as defined in Equation 3 and 4. Unless indicated otherwise, all results use MAPP probabilities, we also compare MAPP and MAPC based results.

Table 7 shows our results for Venezuela election. Here is a description for various acronyms (denoting different techniques) used by us in our results.

| Criteria Name | *Description* |
|---|---|
| *c* | Computing counts based on candidate's full name only. |
| *ca* | Computing counts based on candidate's full name or possible aliases. |
| *cak* | Computing counts based on candidate's full name or simultaneous presence of an alias name as well as one of the keywords in tweet. |
| *cakt* | Limiting every user to a single tweet per day. This criteria is applied on all tweets filtered by criteria *cak*. |

**Table 6: Describing different acronyms used to indicate different techniques in our results.**

Nicolas Maduro won the Venezuela presidential election by a narrow margin of 1.5% Votes. As shown in Table 7, we are able to predict Venezuelan election accurately. However different techniques in using tweets vary in terms of prediction accuracy. Outcome using candidate's full name (denoted by c) are biased towards Nicolas Maduro by 21% compared to Henrique Capriles Radonski. The primary reason of the bias is that most of the tweets referring to Henrique Capriles Radonski donot mention his full name. When we include all aliases while counting relevant tweets (denoted by ca), our bias towards Nicolas Maduro drops from 21% to 2.5%. Searching for keywords, while matching aliases (denoted by cak) further decreases our bias to 1%. Limiting one tweet per day for every user denoted by (cakt) gives our best result by reducing bias to 0.6%.

Table 8 shows results for Paraguay 2013 presidential elections. Horacio Cartes won the election by a margin of 9% over Efrain Alegre. We are able to predict election correctly with an error margin of 10% while using candidate's full name (c). As in the case of Venezuela election, limiting one tweet per day for every user (cakt) gives our best result by reducing error margin to 2%.

| *Candidate* | *Results* | *c* | *ca* | *cak* | *cakt* |
|---|---|---|---|---|---|
| Maduro | 50.6 | 72.0 | 53.0 | 50.5 | 50.3 |
| capriles | 49.1 | 27.0 | 47.0 | 49.5 | 49.7 |

**Table 7: Twitter vote prediction for the 14 April 2013 Venezuela Presidential elections compared with the actual results in percentages.**

| *Candidate* | *Results* | *c* | *ca* | *cak* | *cakt* |
|---|---|---|---|---|---|
| Horacio Cartes | 45.8 | 56.0 | 8.0 | 42.0 | 44.0 |
| Efrain Alegre | 36.9 | 17.0 | 38.0 | 38.0 | 38.0 |
| Mario Fereeiro | 5.8 | 23.0 | 1.0 | 14.0 | 13.0 |
| Anibal Carrilo | 3.3 | 1.0 | 2.0 | 1.0 | 1.0 |

**Table 8: Twitter vote prediction for the 21 April 2013 Paraguay Presidential elections compared with the actual results in percentages.**

Table 9 shows results for Ecuador 2013 presidential elections. Rafael Correa won the election by a margin of 24% over Guillermo Lasso. We are able to predict election correctly with an error margin of 5% while using candidate's full name (c). We didnot have an alias list for Ecuador election.

| *Candidate* | *Results* | *c* | *cakt* |
|---|---|---|---|
| Rafael Correa | 57.0 | 62.0 | 62.0 |
| Guillermo Lasso | 23.0 | 23.0 | 24.0 |
| Lucio Gutierrez | 2.0 | 2.0 | 2.0 |

**Table 9: Twitter vote prediction for the 17 Februaryl 2013 Ecuador Presidential elections compared with the actual results in percentages.**

| *Country* | *c* | *ca* | *cak* | *cakt* |
|---|---|---|---|---|
| Venezuela | 0.22 | 0.02 | 0.00 | 0.00 |
| Paraguay | 0.13 | 0.19 | 0.04 | 0.03 |
| Ecuador | 0.03 | 0.03 | 0.03 | 0.03 |

**Table 10: Root Mean Square Error (RMSE) for all elections**

| *Country* | *c* | *ca* | *cak* | *cakt* |
|---|---|---|---|---|
| Venezuela | 1.00 | 1.00 | 1.00 | 1.00 |
| Paraguay | 0.79 | 0.40 | 0.97 | 0.98 |
| Ecuador | 0.99 | 0.99 | 0.99 | 0.99 |

**Table 11: Correlation Coefficient (CORR) for all elections**

Table 10 shows Root Mean Square Error (RMSE) for all three elections with different criteria as computed in equation 6. Limiting one tweet per day for every user (cakt) gives best results. For Venezuelan election, RMSE improves from 0.22 (c) to 0.00 (cakt). We get an improvement of 0.1 for Paraguay election.

Table 11 shows Correlation Coefficient (CORR) for all three elections with different criteria as computed in equation 7. 'cakt' criteria correlates best with actual results. The improvement in CORR is significant for Paraguay election.

| *Election* | *RMSE* | *CORR* |
|---|---|---|
| Venezuela | 0.00 | 1.00 |
| Paraguay | 0.03 | 0.98 |
| Ecuador | 0.03 | 0.99 |

**Table 12: Showing Mean Square Error (RMSE) and Correlation (CORR) for all three elections using 'cakt' criteria**

Table 12 shows Root Mean Square Error (RMSE) and Correlation (CORR) statistics for all three elections using 'cakt'criteria. We achieve RMSE of less than 0.04 and CORR greater than 0.97 for all three elections.

| *Election* | *MAPP* | *MAPC* |
|---|---|---|
| Venezuela | 0.00 | 0.00 |
| Paraguay | 0.03 | 0.03 |
| Ecuador | 0.03 | 0.05 |

**Table 13: Showing Root Mean Square Error (RMSE) for all three elections with'cakt'criteria using MAPP and MAPC**

Table 13 shows RMSE for all three elections using both MAPP and MAPC. It shows that MAPP criteria performs slightly better than MAPC.

Figure 2 shows the variation of estimated MAPP probabilities of Venezuelan election candidates across one month priot to election day (14 April 2013). The estimated probability of a candidate winning on a day is computed using last week of twitter data as shown in equation 1 and 2. Figure 2 shows that Nicoals Maduro was comfortable ahead of Henrique Capriles Rodanski for most time except in last few days when Henrique Capriles overtook Nicoals Maduro briefly. The time series graph indicates a very close election between both candidates in the last few days, corroborated by the actual results.

Figure 3 shows the temporal variation of Paraguay 2013 election candidate's winning probabilities for one month prior to the election day (21 April 2013). As shown in figure 3, Horacio Cartes lead in estimated probabilities from Twitter for most of the time.

Figure 4 shows the time variation of Ecuador 2013 election candidate's winning probabilities for one month prior to the election day (17 February 2013). Rafael Correa leads in estimated probabilities for all days in prior month and was never overtaken by Guillermo Lasso (candidate who finished second). It indicates a less closer election compared

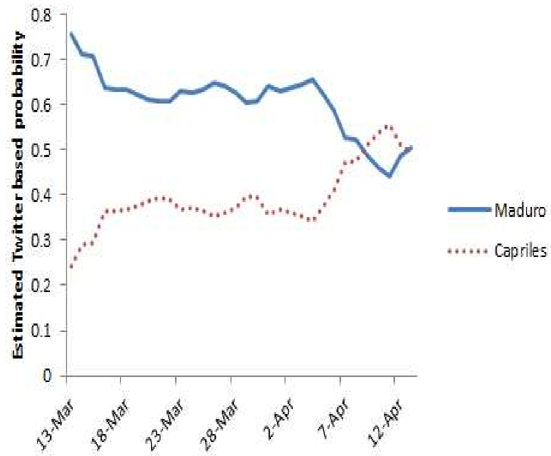


**Figure 2: Shows variation of candidates probabilites of winning the Venezuela election over one month prior to the election. Venezuelan presidential election was held on 14 April 2013. It shows that Nicoals Maduro was comfortably ahead of Henrique Capriles Rodanski for most of the time except in last few days when Henrique Capriles overtook Nicoals Maduro briefly. Maduro regained his lead on last day (12 April) prior to the election. It corroborates with actual very narrow victory of Maduro (1.5%).**

to Venezuela and Paraguay election.

# 5. CONCLUSION

We have collected a large number of global Twitter messages (approximately 13 billion) and showed how they can be used for predicting the results of the Venezuelan, Paraguayan and Ecuadorian Presidential election of 2013. Our findings suggest that while there is a strong correspondence between the share of tweets and the share of votes in the elections, counting the tweets that mention a candidate's conventional name is not sufficient to obtain good predictions. We find that using appropriate aliases for a candidate can greatly enhance the estimate of the mention count of relevant tweets to an election, but unconstrained use of aliases might also negatively affect the predictive power of tweets.

Furthermore, we show that using appropriate keywords in conjuction with aliases while searching for political tweets strengthens the predictive power of tweet based system. We tested the effects of improving our Twitter based system by ignoring multiple tweets from a single user. Finally, we show that computing a Moving Average Aggregate Probability (MAPP) of a candidate over a period of 7 days results in reasonably accurate election predictions. The Average Root Mean Square Error (RMSE) of our final system was less than 0.03 (see Table 10).

We hope to improve these results in the future, building on the knowledge we have obtained in this study. We will weight the mentioned counts based on the average number of tweets by a user, use different weights for keywords and then assess its impact on election prediction accuracy.

# 6. ACKNOWLEDGEMENTS

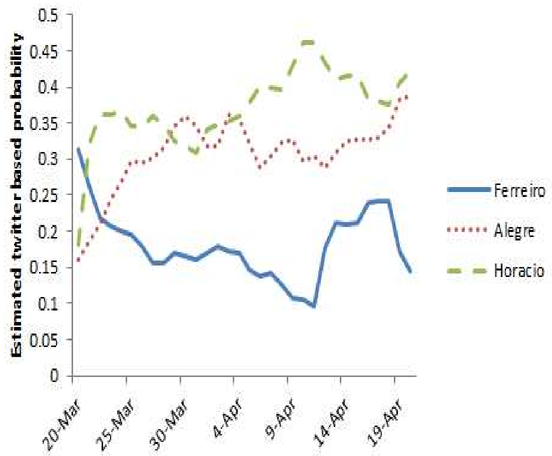


**Figure 3: Shows variation of candidate probabilites of winning the Paraguay election over one month prior to the election. Paraguayan presidential election was held on 21 April 2013. It shows that eventual winner Horacio Cartes was comfortably ahead of Efrain Alegre for most of the time. His lead was much narrower in the last few days leading to the election. It corroborates with actual somewhat narrow victory of Cartes (9%).**

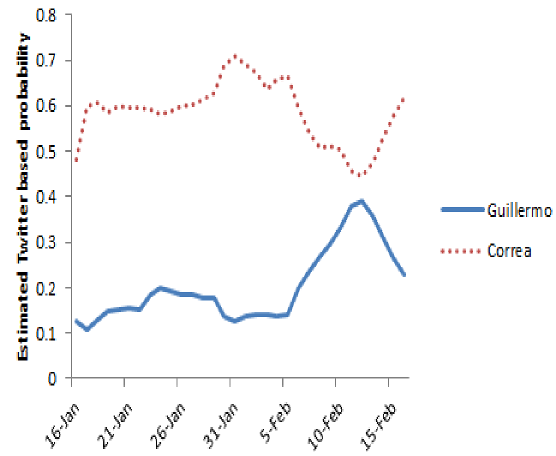


**Figure 4: Shows variation of candidate probabilites of winning the Ecuadorian 2013 election over one month prior to the election. The election was held on 17 February 2013. It shows that Rafael Correa was comfortably ahead of Guillermo Lasso for almost all last month. It corroborates with actual comfortable margin of victory of Correa (34%).**

We thank Saurabh Khanwalkar for helpful guidance regarding the revision. We acknowledge helpful discussions and valuable suggestions from Dave Stallard, Plamen Petrov and Saurabh Khanwalkar.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contact number D12PC00347.

The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

## 7. ADDITIONAL AUTHORS

## 8. REFERENCES

[1] The history of Twitter. http://www.socialnomics.net/2013/01/23/the-history-of-twitter/

[2] Twitter Decahose. http://gnip.com/twitter/decahose/

[3] JSON in Java. http://json.org/java/

[4] Lunden, Ingrid Twitter May Have 500M+ Users But Only 170M Are Active, 75TechCrunch. Retrieved 2 February 2013

[5] Johan Bollen, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), Pages1-8, 2011.

[6] Sitaram Asur and Bernardo A. Huberman. Predicting the future with socialmedia. 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, Washington, DC, USA, 2010. IEEE Computer Society.

[7] V. Lampos and N. Cristianini. Tracking the ïňĆu pandemic by monitoring the social web In Cognitive Information Processing (CIP), 2010 2nd International Workshop on, pages 411 - 416, june 2010

[8] A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pages 178 - 185, 2011

[9] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010.

[10] A. Tumasjan, T. Sprenger, P. G. Sandner and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proc. of 4th ICWSM, 178 - 185. AAAI Press.

[11] R. Cozma and K. Chen. Congressional Candidates'Use of Twitter During the 2010 Midterm Elections: A Wasted Opportunity? 61st Annual Conference of the International Communication Association, 2011.

[12] E. Tjong, K. Sang and J. Bos. Predicting the 2011 Dutch Senate Election Results with Twitter. In Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks.

[13] D. Gayo-Avello. I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper - A Balanced Survey on Election Prediction using Twitter Data. http://arxiv.org/abs/1204.6441v1

[14] M. Skoric, N. Poor, P. Achananuparp, E-P. Lim, and J. Jiang. Tweets and Votes: A Study of the 2011 Singapore General Election. In Proceedings of the 45th Hawaii International Conference on System Sciences, 2012

[15] Andreas Jugherr, Pascal Jurgens and Harald Schoen. Why the pirate party won the german election of 2009 or the trouble with predictions: A response to a. Tumasjan, T. O. Sprenger, P. G. Sander and I. M. & welpe ‘predicting elections with twitter: What 140 characters reveal about political sentiment’. Social Science Computer Review, 2011

[16] A. Fonseca. Modeling Political Opinion Dynamics Through Social Media and Multi-Agent Simulation. First Doctoral Workshop for Complexity Sciences.

[17] D. Gayo-Avello, P. T. Metaxas and E. Mustafaraj. Limits of Electoral Predictions using Social Media Data. In Proceedings of the ICWSM 2011, 2011.

[18] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini, F. Menczer. Detecting and Tracking Political Abuse in Social Media. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media ICWSM 2011.