FORECASTING PERUVIAN PRESIDENTIAL ELECTION 2016 USING SUPERVISED AGGREGATED SENTIMENT ANALYSIS

Florencio Paucar Sedano

Dipartimento di Informatica
Corso di Laurea Magistrale in Informatica per l'economia e per l'azienda
(Business Informatics)

2nd December 2016

Outline

Introduction

Data Preparation

Modelling

Evaluation

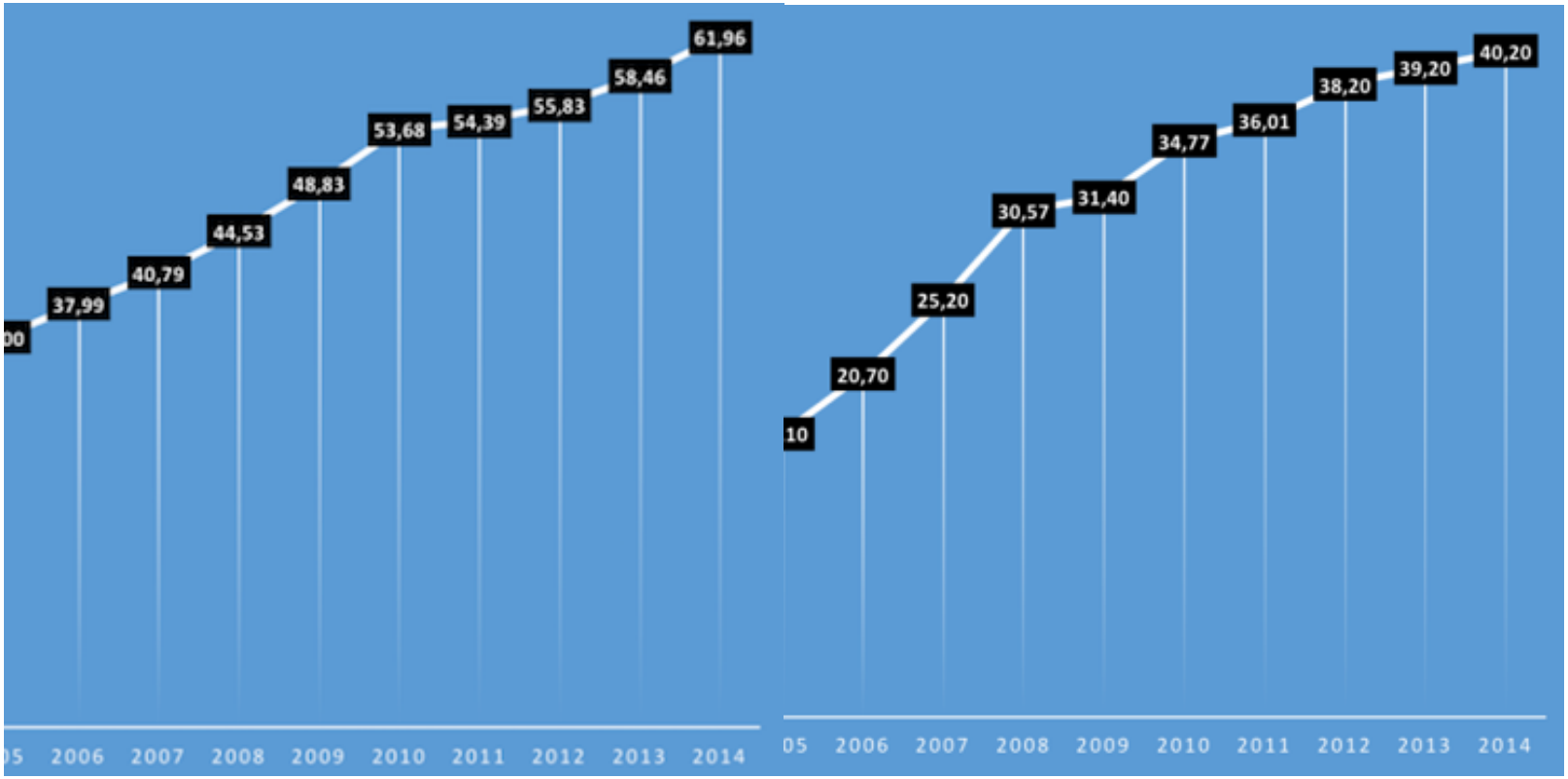
Conclusions

Introduction

- Discusses the political "sentiment" of the twitter information produced from 11 April to 05 June, during the second round Peruvian presidential election between Ms. Keiko Fujimori from Fuerza Popular and Mr. Pedro Pablo Kuczynski from Peruanos por el Kambio political party using the supervised aggregated sentiment analysis method (SASA).
- Considers past electoral analysis realized in United States, France, and Italy. The estimation results with an average mean absolute error of approximately 2.5 points respect to the official results.

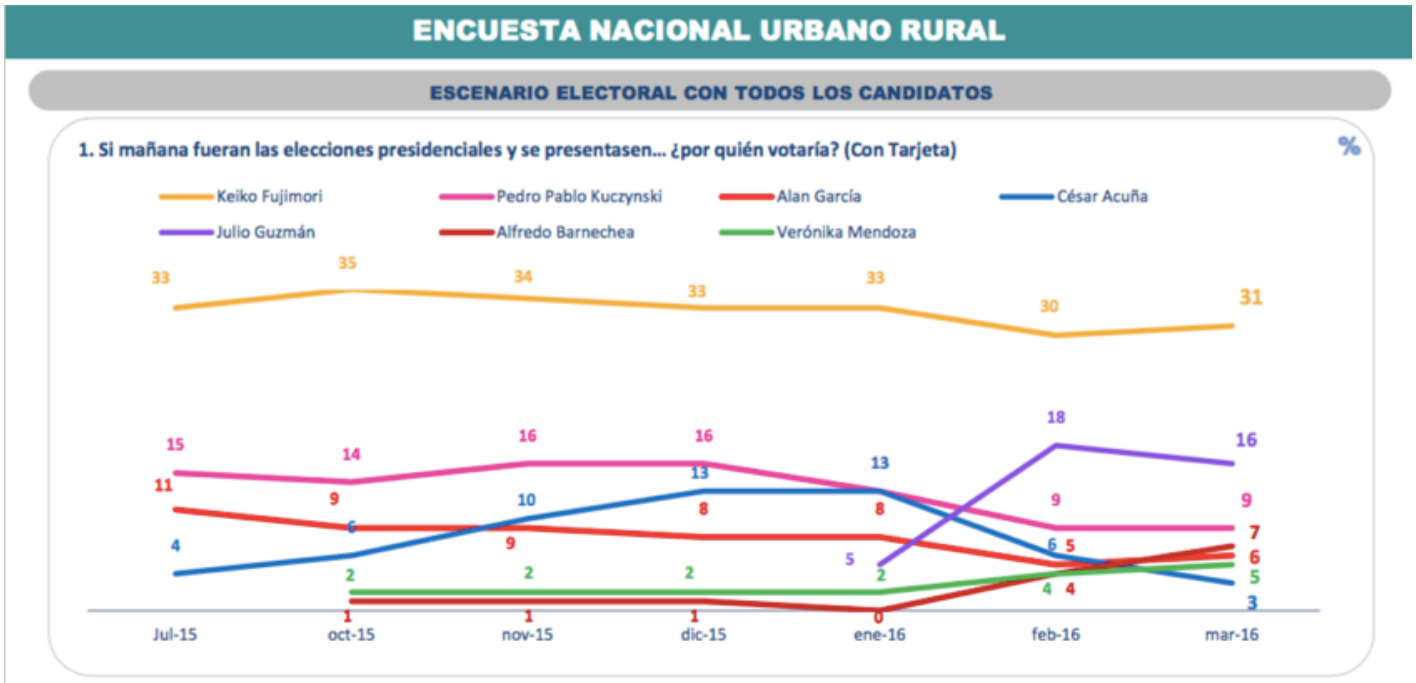
Motivation 1

- Measure the effectivity of the SASA method in emerging internet communities such as Peru.
  E.g. Number internet users Italy - Peru (The World Bank).



Motivation 2

- Special case of the candidate Julio Guzman who obtained a high popularity in the first round of the Peruvian presidential election using social networks (Fig. Ipsos Polling Company).



Objectives

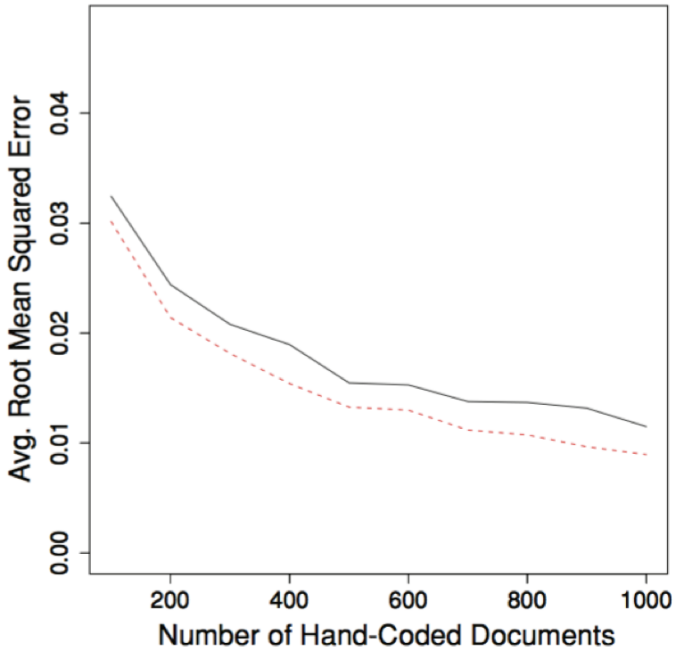
- Analyze the polls behaviors during the Peruvian election race.
- Predict the final results.
- Verified the accuracy of the supervised aggregated sentiment analysis method for emerging social networks communities such as Peru, with approximately 4 million of twitter users and 22 million voters.

Data Preparation

1. Collection using political party and candidates names keywords.
2. Separation of tweets wrote in languages different to Spanish and retweets. Finally, obtaining 302 105 tweets to be analyzed.
3. Text processing of the twitter data (tranformation in lower case, elimination of Urls, removing of punctuation characters, tokenization of text)

Hand codding

Consist in reading and assign a certain class to a subsample of the twitter data collected. This subsample is the training set which will be used by the ReadMe algorithm to classify the test set (D. Hopkins G. K., 2010).



Aggregated Statistical Analysis

Uses ReadMe algorithm to obtain the proportion of opinions expressed in the entire dataset related to the categories defined using aggregated classification since individual document classification can lead to biased estimates of the document category proportions. Variables of the algorithm:

- **S as words used in the texts.**
- **D represents opinion of people expressed in the tweets.**
- **K is the number of stems kept in the stemming phase.**
- **J is number of categories to be considered in the analysis.**

Readme Algorithm

The objective is estimate P(D) = P(D/S) P(S)..... (1)

- **P(S) is a vector of dimension** $2^K x1$ **and represents the tabulation of frequencies of word profiles in the whole population of texts**. If k=3 word stems, P(S) would contains $2^3 = 8$ patterns occurring in the next set of documents: 000, 001, 010, 011,100,101,110, 111.

- **P(D/S) is a matrix of dimension** $2^K xJ$ **and estimates the conditional distribution of word profiles within the training set.**

- **P(D) is a vector of dimension** $Jx1$**, which is the J-vector quantity of interest.**

Readme Algorithm

The first approach to obtain P(D) uses any individual classifier. Unfortunately, the estimates of P(D/S) in the training set are biased and presents high variability due to the noise in the tweets. Thus, Readme focuses in :

$P(S) = P(S/D)P(D)$.....(2)

P(S/D) is not observed on the whole data set and it is estimated by hand-coding of the training set.

$P^h(S/D) = P(S/D)$.....(3)

By solving the equation (2) via standard regression algebra:

P(D), unknown regression coefficients which could be called β.

P(S/D), explanatory variables which could be called matrix X.

P(S), dependent variable which could be called Y.

Readme Algorithm

The equation (2) become Y = X β (with no error term). The result P(D) is calculated via usual regression or via standard constrained least squared to ensure that elements P(D) are each in [0,1] and collectively sum to 1.
The equation (2) could be written in the following way.

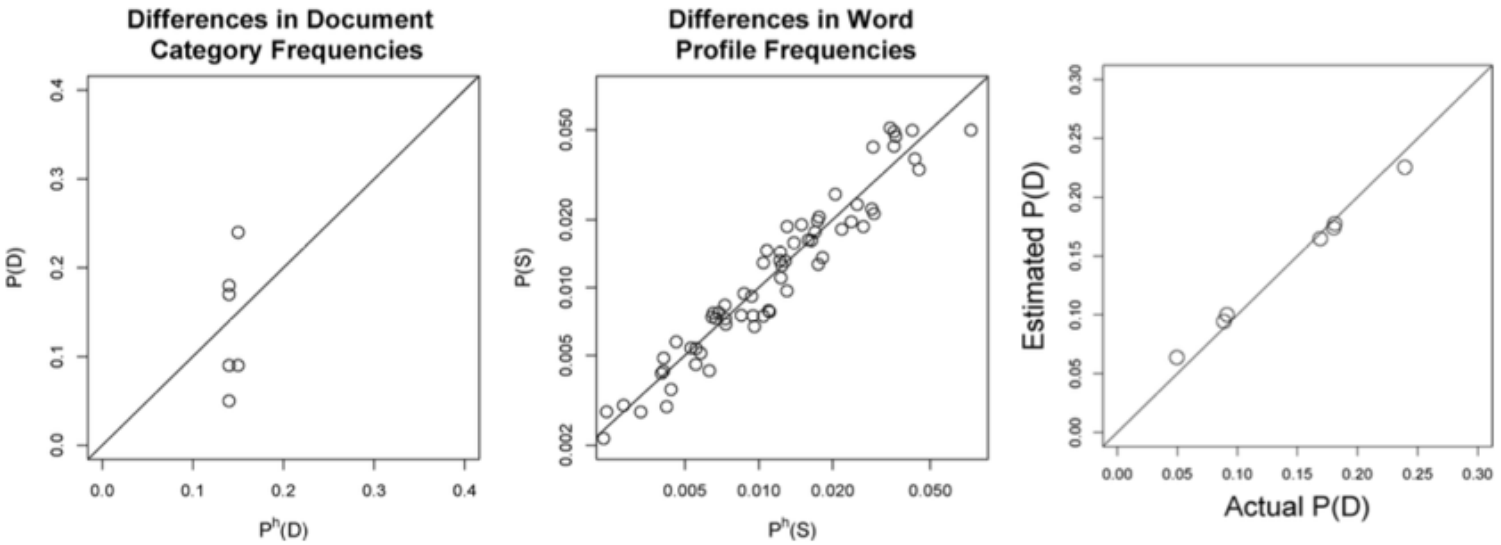
$$P(D) = P(S/D)^1 P(S) \text{.....(4)}$$

Finally, using the equation (3) and (4), it is obtained the objective of the readme algorithm:

$$P(D) = P^h(S/D)^{-1} P(S) \text{.....(5)}$$

Readme Algorithm

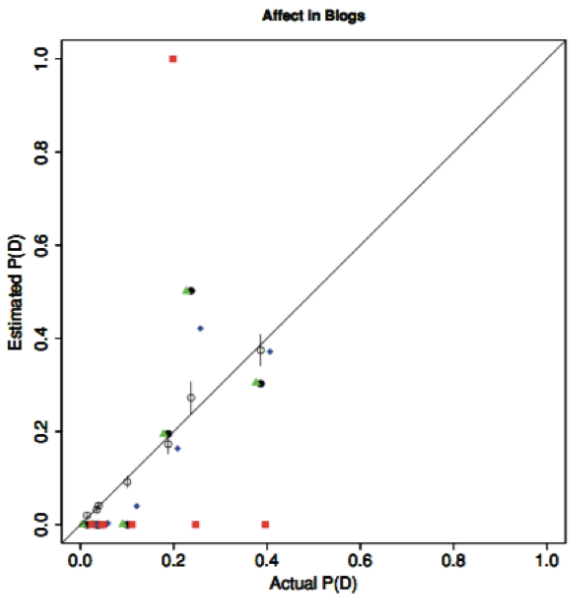
No biases would be produced if a word become more popular between the time when the training set was hand coded and the population documents were collected. Likewise, if documents in certain categories are more prevalent in the population than training set, no biases would be present (D. Hopkins G. K., 2010).



Notes: For both $P(D)$ on the left and $P(S)$ in the center, the distributions differ considerably. The direct sampling estimator, $P^h(D)$, is therefore highly biased. Yet, the right panel shows that our nonparametric estimator remains unbiased.

Readme Algorithm

An empirical evidence of the performance of the ReadMe aproach in comparison with individual classificators using Support Vector Machine on data extracted from blog posts (D. Hopkins, 2010).



Notes: The plot gives the estimated document category frequencies (vertically) by the actual frequencies (horizontally). Our nonparametric approach is represented with black open circles, with 95% confidence intervals as vertical lines. Aggregated optimized SVM analyses also appear for radial basis (black dots), linear (green triangles), polynomial (blue diamonds), and sigmoid kernels (red squares). Estimates closer to the 45° line are more accurate.

Hand coding parameters calibration

Assignation of specific categories to the tweets:
Category 0: Tweets which do not express leaning in favor of any candidate or present unrelated information to the Peruvian election.
Category 1: This category expresses an explicit support to the candidate Mr. Kuczynski.
Category 2: This category exhibits a favor to the candidate Mrs. Fujimori.
Consider as real intention to cast a vote in favor of a specific candidate:

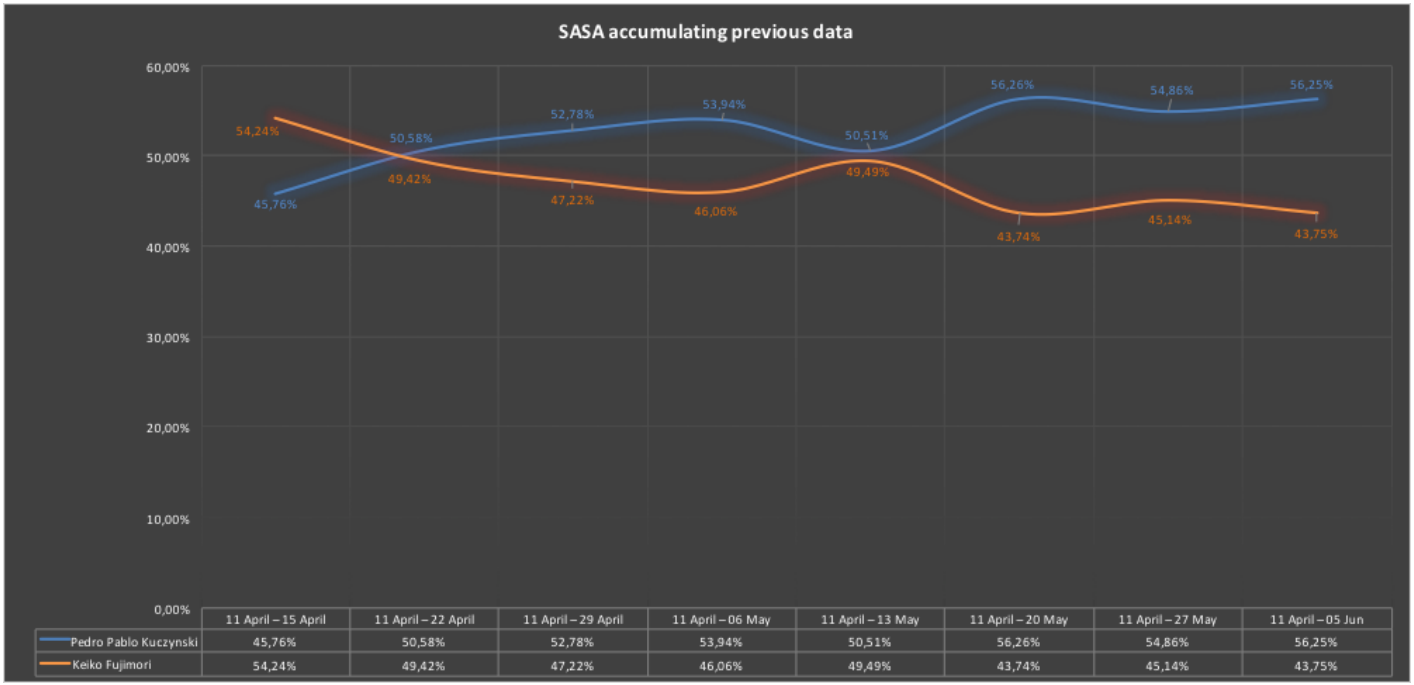
- An explicit statement to vote for a candidate.
- Contains a statement in favor of a specific candidate together with a message or a hashtag.
- Presents a negative statement opposing a candidate with a message or a hashtag related to the rival candidate.

Model Building

Two important aspect were analyzed to build the models:

- Time windows.
- Period hand coded data

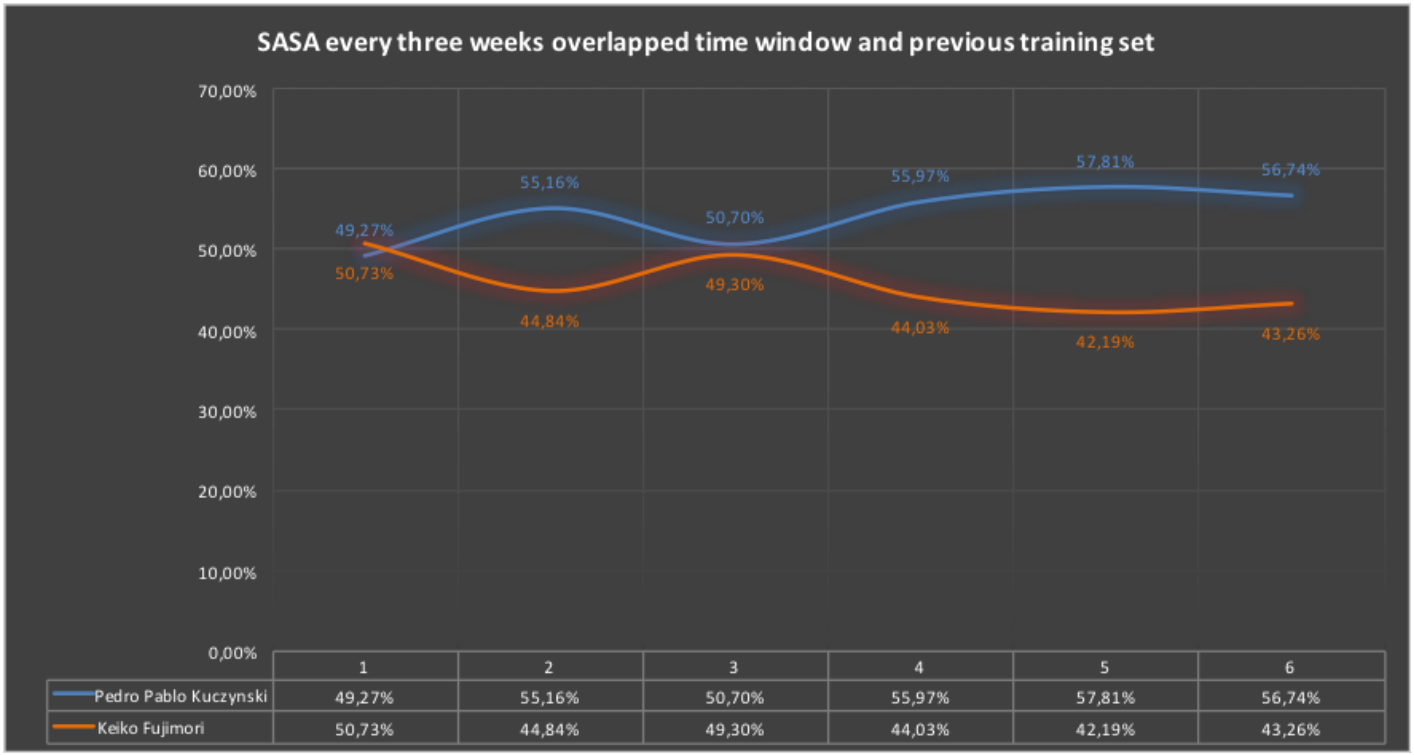
Representative models: Model 1



	11 April – 15 April	11 April – 22 April	11 April – 29 April	11 April – 06 May	11 April – 13 May	11 April – 20 May	11 April – 27 May	11 April – 05 Jun
Pedro Pablo Kuczynski	45,76%	50,58%	52,78%	53,94%	50,51%	56,26%	54,86%	56,25%
Keiko Fujimori	54,24%	49,42%	47,22%	46,06%	49,49%	43,74%	45,14%	43,75%

Model Building

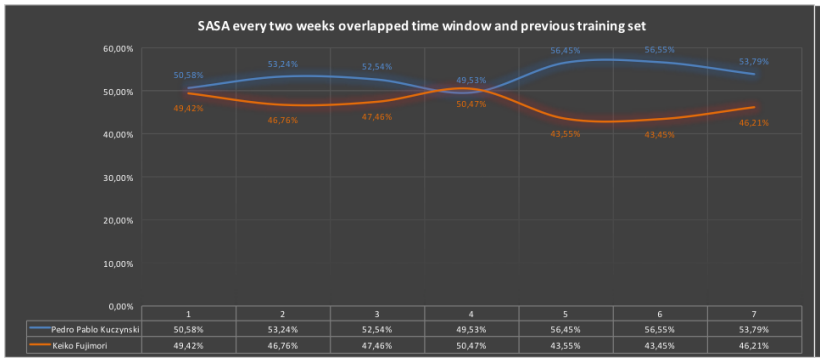
Model 2:

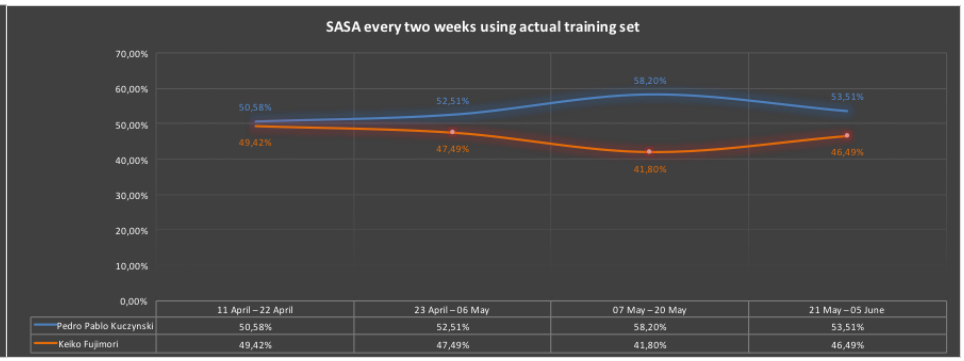


	1	2	3	4	5	6
Pedro Pablo Kuczynski	49,27%	55,16%	50,70%	55,97%	57,81%	56,74%
Keiko Fujimori	50,73%	44,84%	49,30%	44,03%	42,19%	43,26%

Model Building

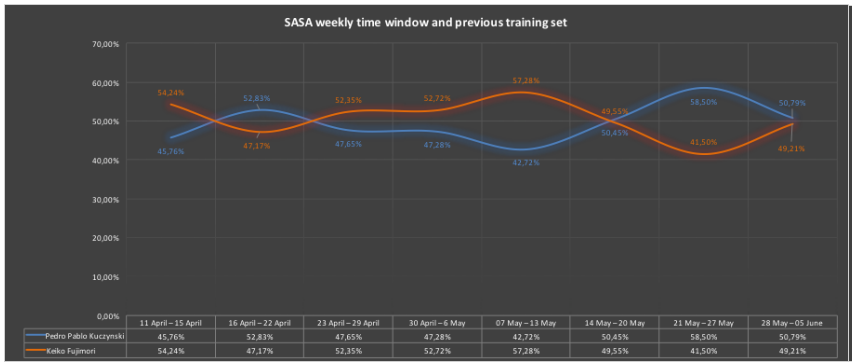
Model 3 and 4:

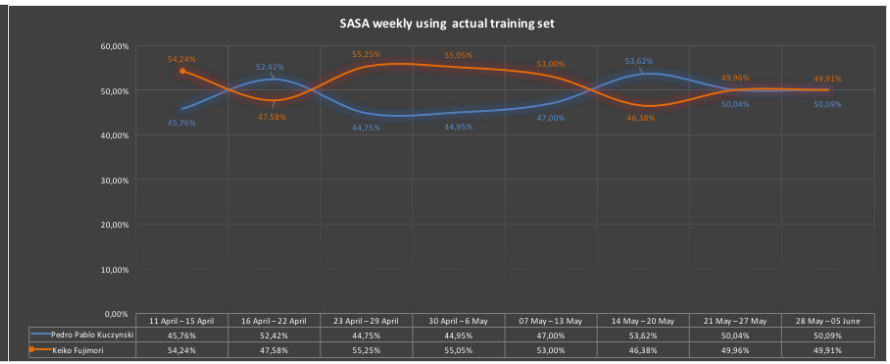




Model Building

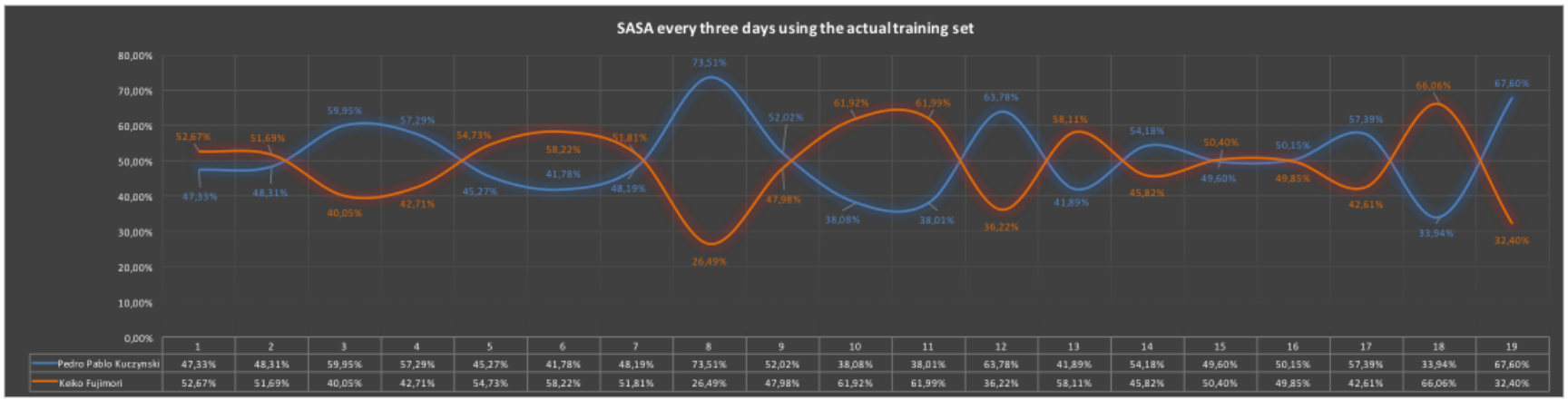
Model 5 and 6:





Model Building

Model 7:

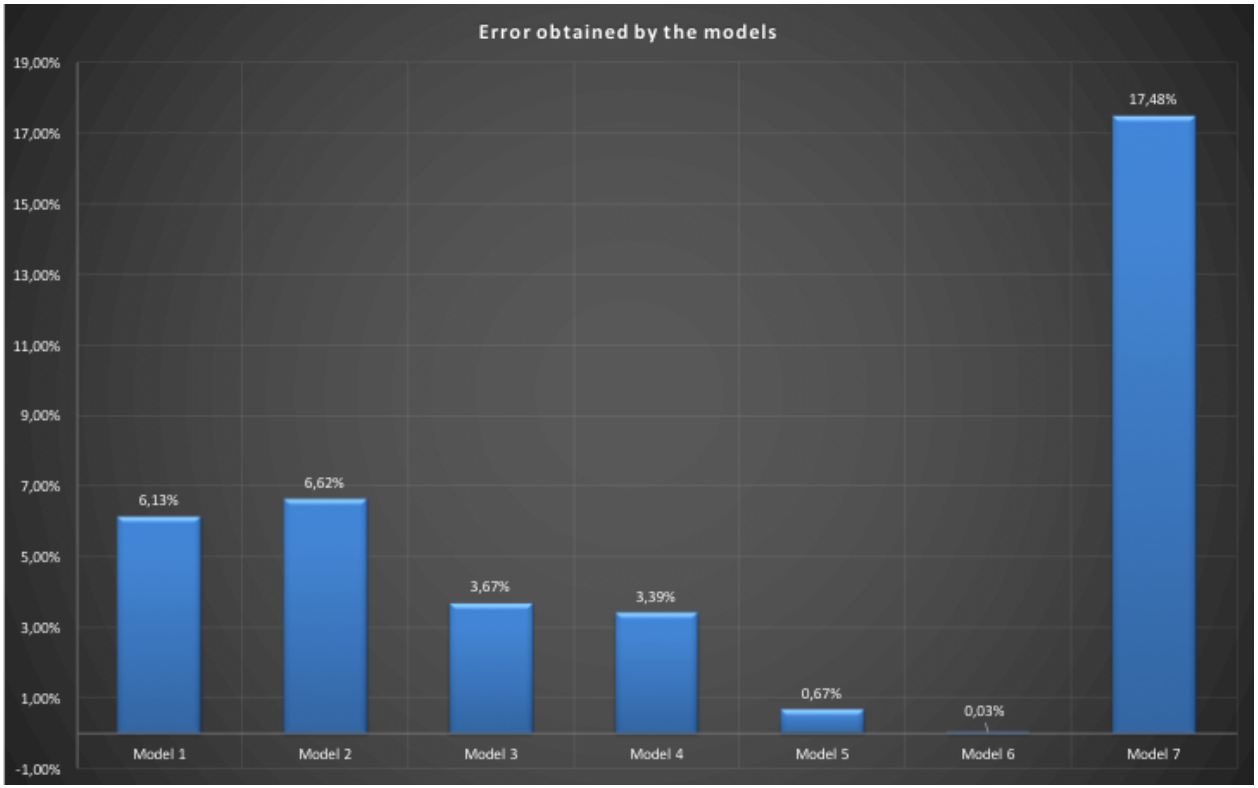


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Pedro Pablo Kuczynski	47,33%	48,31%	59,95%	57,29%	45,27%	41,78%	48,19%	73,51%	52,02%	38,08%	38,01%	63,78%	41,89%	54,18%	49,60%	50,15%	57,39%	33,94%	67,60%
Keiko Fujimori	52,67%	51,69%	40,05%	42,71%	54,73%	58,22%	51,81%	26,49%	47,98%	61,92%	61,99%	36,22%	58,11%	45,82%	50,40%	49,85%	42,61%	66,06%	32,40%

Model assessment

Based in two aspects:

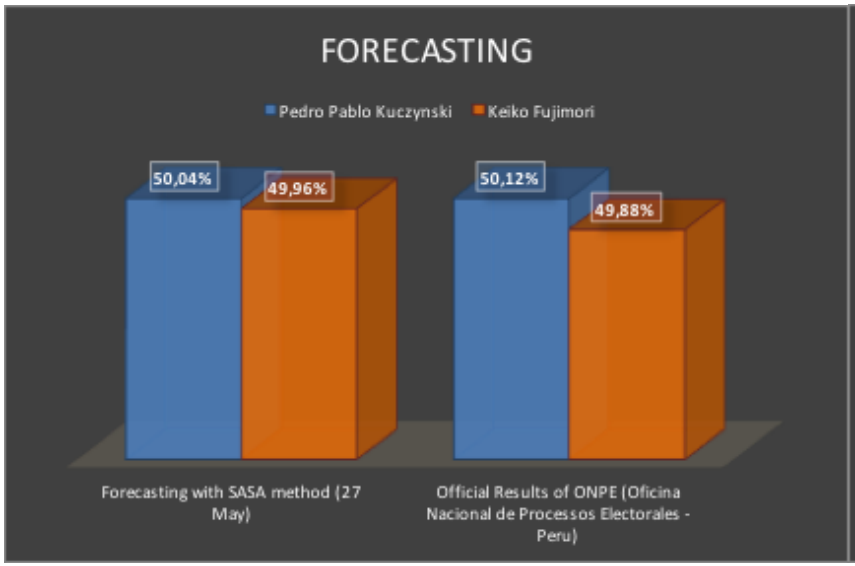
1. Behaviors of the polls (A. Ceron L. C., 2014b).
2. Accuracy of the final results (Post electoral comparison)

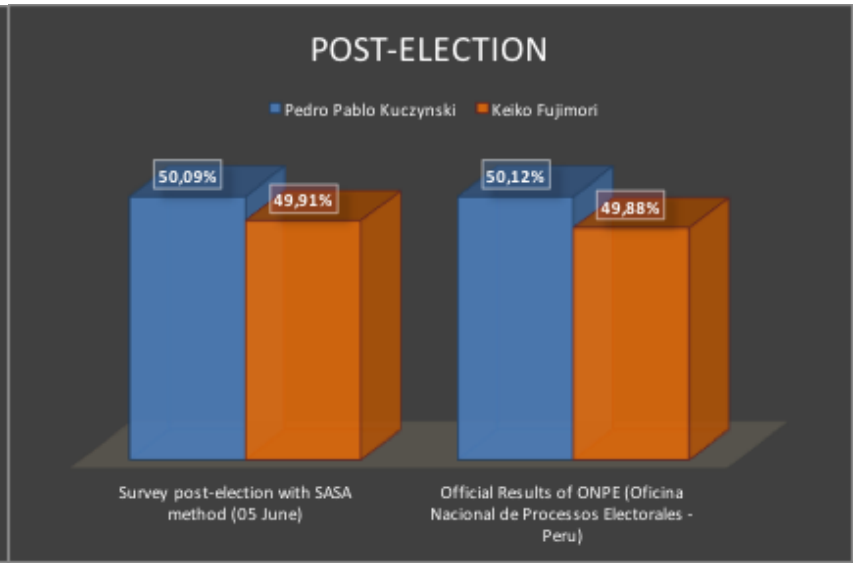


Evaluation

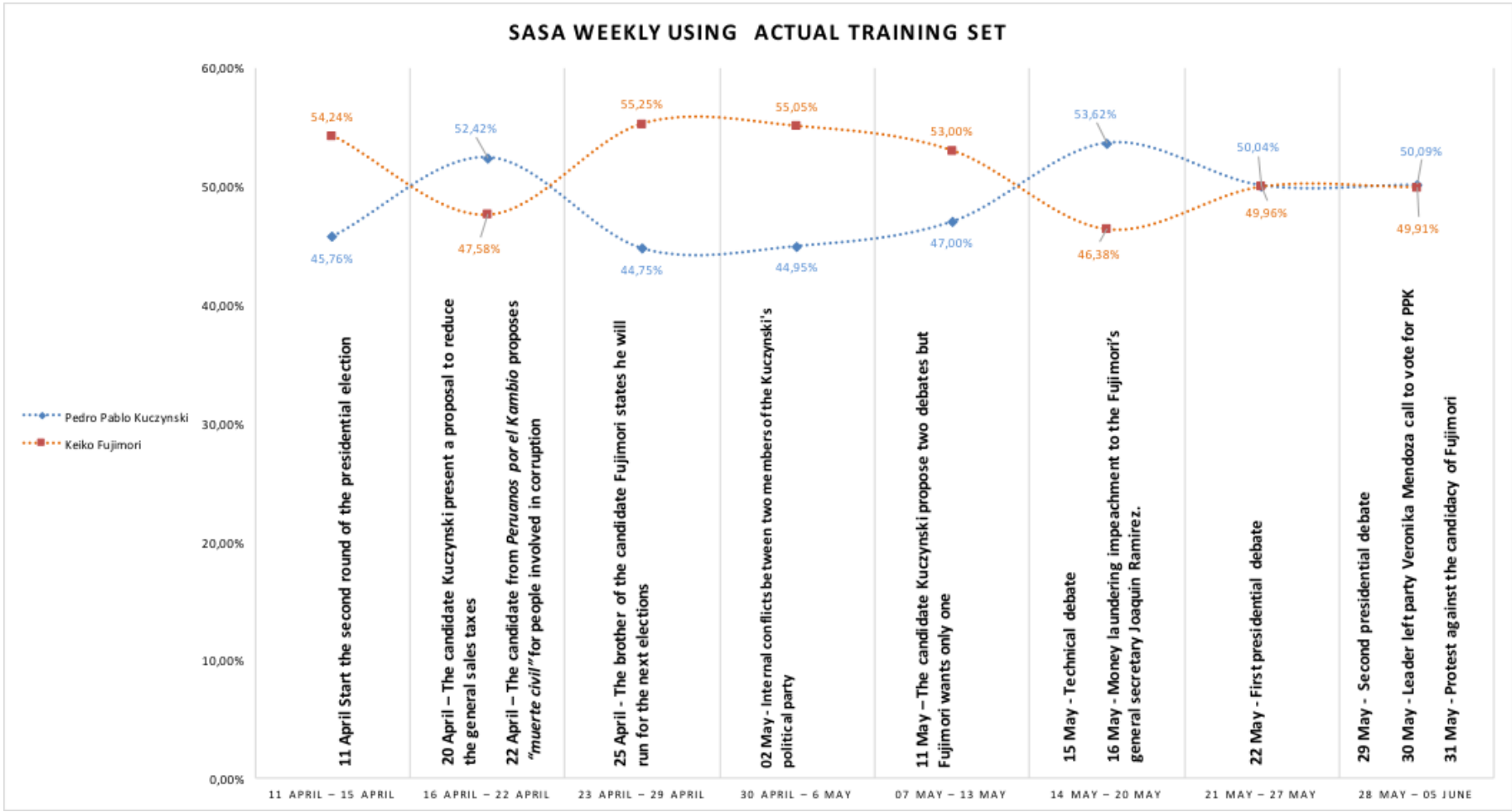
The evaluation of the performance of the selected model 6 follow the next criteria:

1. Comparison of the results of the model with official results.
2. Analysis of the main events that happened during the electoral campaign.
3. Comparison of the selected model with traditional Mass-Surveys.





Analysis of the main events during the electoral campaign



Comparison of the selected model with traditional Mass-Surveys

The following official polling firms who created off-line surveys are considered in these analysis:

- Compania Peruana de estudios de Mercado y Opinion Publica.
- IPSOS Peru.
- DATUM internacional.
- GFK Peru.

All polling firms gave a wrong prediction one week before the presidential election and also all the post-election off-line surveys presented more margin of error than the SASA method.

Conclusions

- The SASA method obtained accurate results since it is focuses on the estimation of aggregated distribution of opinions.
- The appropriate model to forecasts elections in similar contexts contemplates the use of weekly time windows analysis and the use of hand coded data from the same period.
- The comparison between the created model and some traditional mass-surveys permits to know its good performance.
- The analysis of the behavior of the polls assents to correlate some important events with the performance of the candidates on the polls.
- The model created is appropriate and consent to verify the accuracy of the SASA method for emerging social networks communities such as Peru.