

RESEARCH

Supplementary Information for: Why polls fail to predict elections

Zhenkun Zhou¹, Matteo Serafino², Luciano Cohan³, Guido Caldarelli^{4,5,6,7} and Hernán A. Makse^{8*}

*Correspondence:

hmakse@ccny.cuny.edu

⁸Levich Institute and Physics Department, City College of New York, 10031 New York, USA
Full list of author information is available at the end of the article

Author details

¹School of Statistics, Capital University of Economics and Business, 100070 Beijing, China. ²IMT School for Advanced Studies, 55100 Lucca, Italy. ³Seido, Buenos Aires, Argentina. ⁴Department of Molecular Sciences and Nanosystems, Ca' Foscari University of Venice, 30172 Venice, Italy. ⁵European Centre for Living Technology, Italy Venice, Italy. ⁶Institute for Complex Systems, Consiglio Nazionale delle Ricerche, UoS Sapienza, 00185 Rome, Italy. ⁷London Institute for Mathematical Sciences, W1K2XF London, United Kingdom. ⁸Levich Institute and Physics Department, City College of New York, 10031 New York, USA.

Tweets	Bots	Users
Total	538359	67336507
Daily	2243	280568
Daily classified	1617	211229
Daily classified MP	619	114653
Daily classified FF	998	96576

Table S1: Tweet statistics. We report the total number of tweets collected, the average daily number of tweets classified (Daily classified), and the average daily classified tweets for each candidate.

Users	Bots	Users
Total	17953	2252551
Daily	732	83330
Daily classified	560	63808
Daily classified MP	198	31497
Daily classified FF	362	32310

Table S2: User statistics. We report the total number of users collected, the average daily number of users classified (Daily classified), and the average daily classified users for each candidate.

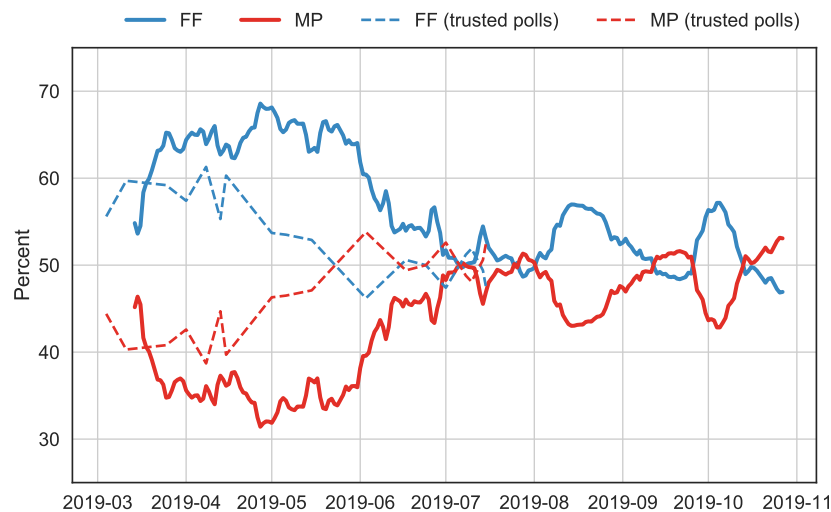


Fig. S1: Instantaneous prediction compared with trusted polls. Thick lines represent AI prediction and dashed lines represent trusted polls. We use the 7, 14, 21, 28-day windows for the instantaneous opinion model. We calculate the Pearson's correlation coefficient between MP from results of instantaneous models with different window sizes and that from pollsters. The coefficients are respectively 0.372, 0.464 (14-day window, the highest), 0.408, and 0.438. Therefore, we select 14-days windows for the instantaneous opinion model.

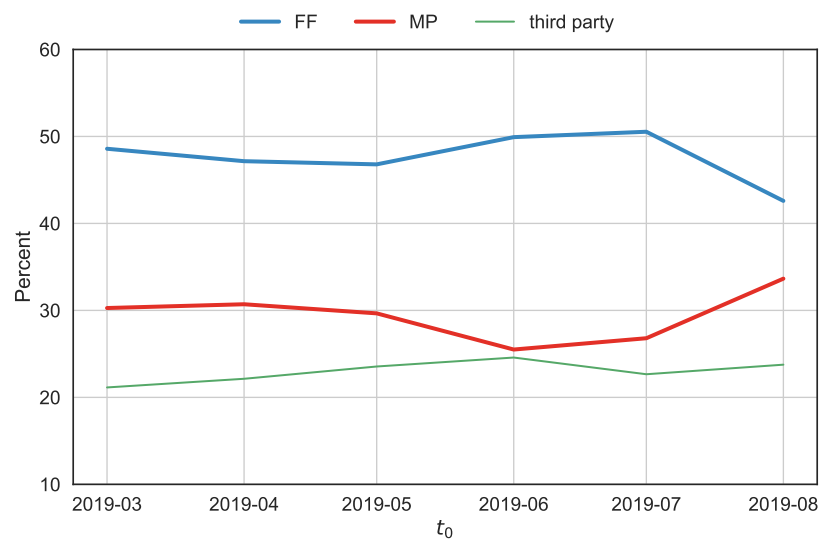


Fig. S2: Cumulative prediction with T_0 =March 1 (2019-03) until August 1 (2019-08) before PASO. As for the cumulative opinion model, we select the initial date according to the mean absolute error (MAE) between the predicted and official results on the primary election, which can be seen in Fig. 2, and the errors are respectively 1.00 (the least, Mar), 1.13, 2.07, 4.20, 3.33 and 3.50 (T_0 =Mar, Apr, May, Jun, Jul, Aug 1). Therefore, we select March 1 as the initial date with the least error.

#	Hashtag	Camp	Count
1	sisepuede	M	216757
2	macri	M	102172
3	albertopresidente	K	65644
4	axelgobernador	K	60920
5	juntosporelcambio	M	52617
6	yovotomm	M	49557
7	ladamosvuelta	M	47546
8	cfk	K	38718
9	cambemos	M	38538
10	sevan	K	32110
11	habraconsecuencias	K	28406
12	24a	M	23819
13	macrihacete cargo	K	21364
14	novuelvenmas	M	21054
15	cronicaanunciada	K	20293
16	frentedetodos	K	18115
17	albertoycristina	K	17482
18	sinceramente	K	15403
19	juntosomosimparables	M	14405
20	sevanenprimeravuelta	K	14091

Table S3: The top 20 hashtags from March and July in 2019. The camp field represents the classification: M stays for Macri and K for Fernandez (from the name of the running mate Cristina Fernandez de Kirchner). Count indicates the number of times a given hashtag appears in the dataset.

#	Hashtag	Camp	Count
1	cambiemos	M	150754
2	macri	M	120363
3	cfk	K	63052
4	sinceramente	K	56198
5	habraconsecuencias	K	52896
6	novuelvenmas	M	39352
7	juntosporelcambio	M	35186
8	macritevuelvoelegir	M	32241
9	cronicaanunciada	K	30954
10	massa	K	29565
11	axelgobernador	K	29039
12	navarro2019	K	27885
13	defensoresdelcambio	M	27010
14	fuerzacristina	K	21342
15	andatemacri	K	19486
16	chaumacri	K	19184
17	debodecir	M	19173
18	hayotrocaminio	K	19152
19	mm2019	M	16117
20	sracristinalecuentoque	M	16082

Table S4: The top 20 hashtags from August and October in 2019. The camp field represents the classification: M stays for Macri and K for Fernandez (from the name of the running mate Cristina Fernandez de Kirchner). Count indicates the number of times a given hashtag appears in the dataset.

Introduction of five trusted pollsters

- **Real-Time Data** is an innovative consultancy firm based in Miami that provides systematic, permanent, and in-detail public opinion data to local and foreign investors, researchers, and decision-makers with interests in Argentina and the rest of Latin America. Official site: <https://inrtd.com>.
- **Management Fit** focuses on the public opinion research of the political and social situation, in order to provide reliable results, design, and management of strategies that allow capitalizing decision-making. Official site: <https://myfconsultora.com.ar/>.
- **Opinaia**, since its beginning in 2010, is constituted with the aim of addressing, with a technological view, the challenges the new stage of market research requires. They achieving an ongoing integration between research and technology to the benefit of new market challenges. Official site: <https://www.opinaia.com/>.
- **Giacobbe** conducts studies and research for national and international clients since 1992. they have consolidated our experience in areas of government and politics, civil associations, media, and market research. Official site: <https://giacobbeconsultores.com/>.

- **Elypsis** is a trusted polling firm in Buenos Aires, Argentina, that focuses on the public opinion area. Elypsis employs a combination of IVR of landline numbers complemented with online opt-in samples to obtain a public opinion in Argentina. Official site: <https://www.elypsisweb.com/>.