Zhou *et al. J Big Data* (2021) 8:137
https://doi.org/10.1186/s40537-021-00525-8

Journal of Big Data

RESEARCH Open Access

Why polls fail to predict elections

Zhenkun Zhou[1], Matteo Serafino[2], Luciano Cohan[3], Guido Caldarelli[4,5,6,7] and Hernán A. Makse[8*]

*Correspondence:
hmakse@ccny.cuny.edu
[8] Levich 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

Abstract

In the past decade we have witnessed the failure of traditional polls in predicting presidential election outcomes across the world. To understand the reasons behind these failures we analyze the raw data of a trusted pollster which failed to predict, along with the rest of the pollsters, the surprising 2019 presidential election in Argentina. Analysis of the raw and re-weighted data from longitudinal surveys performed before and after the elections reveals clear biases related to mis-representation of the population and, most importantly, to social-desirability biases, i.e., the tendency of respondents to hide their intention to vote for controversial candidates. We propose an opinion tracking method based on machine learning models and big-data analytics from social networks that overcomes the limits of traditional polls. This method includes three prediction models based on the loyalty classes of users to candidates, homophily measures and re-weighting scenarios. The model achieves accurate results in the 2019 Argentina elections predicting the overwhelming victory of the candidate Alberto Fernández over the incumbent president Mauricio Macri, while none of the traditional pollsters was able to predict the large gap between them. Beyond predicting political elections, the framework we propose is more general and can be used to discover trends in society, for instance, what people think about economics, education or climate change.

Keywords: Election prediction, Social-desirability biases, Social networks, Big-data analytics, Opinion analysis, Machine learning, Natural language processing, Artificial Intelligence

Introduction

Traditional polling methods using random digit dial phone interviews, opt-in samples of online surveys and interactive voice response are failing to predict election outcomes across the world [1–4]. The failure of traditional surveys has also been widely discussed in the press [5] and on the specialized literature [2]. For instance, the victory of Donald Trump in the US 2016 presidential election came as a shock to many, as none of the pollsters and political journalists and pundits, including those in Trump's campaign, could predict this victory [6].

The reasons for the failure of pollsters to predict elections are believed to be many [2, 6]. Firstly, the percentage of responses to traditionally conducted surveys has decreased and it is becoming increasingly difficult to get people's opinion. Response rates in telephone polls with live interviewers continue to decline, and it has reached a 6% lower limit recently [7]. Response rates could be even lower for other methodologies, like internet polling or interactive voice response. Compounded with declining

response rates is the concomitant problem of mis-representation of the survey samples. That is, the sample surveyed by pollsters does not represent the demographic distributions of the general population. This problem is ameliorated by re-weighting the surveys sample according to the general demographics of the population in a process called sample-balancing or raking [8, 9]. However, in countries where the vote is not obligatory, re-weighting a sample to the general population demographics (obtained from Census Bureau [1, 2]) could fail since the general population does not match necessarily the demographics of the voter turnout. Thus, it is important to predict which demographic groups will turn out at the voting station. For instance, if an underrepresented group in the polls, "the hidden vote", decides to vote on election date, the re-weighting fails leading to inaccurate results. The issue is believed to be one of the major reasons for the generalized failure of pollsters to predict the triumph of Trump in the 2016 US presidential election, where groups generally defined as "white voters without college degree" mostly voted for Trump but were undersampled by pollsters. Even with this historical information at hand, which supposedly allowed pollsters to resample their surveys more carefully, pollsters again under-predicted the support for Trump in the subsequent 2020 presidential election in some states or under-predicted the voter turnout supporting Biden in newly created battleground states [10]. The inability to accurately predict the voter turnout to deal with sampling mis-representation might render the pollsters obsolete.

While there is increasing evidence [2, 7] that nonresponse and mis-representation bias might be the reason that polls are not producing accurately matched election results, these may not be the only problem of traditional methods of polling. Traditional surveys in heavily polarized campaigns are affected by social-desirability biases (also called Bradley effect) [11, 12]. For instance, the tendency of survey respondents not to tell the truth of intention of support for controversial candidates which could open themselves to social ostracism. We show below, that this was a major reason for the pollster's failure to predict the 2019 Argentina election, which involved a controversial candidate (Cristina Fernández) who was heavily under-predicted in traditional polls. All these peculiarities together make them difficult for traditional polls to correctly predict the results of elections.

Monitoring social networks [13, 14] represents an alternative to polls to capturing people's opinions since it overcomes the low-response rate problem and it is less susceptible to social-desirability biases [11, 12]. Indeed, social network users continuously express their political preferences in online discussions without being exposed to direct questions. One of the most studied social networks is the microblogging platform Twitter [15–18]. Twitter-based studies generally consist of three main steps: data collection, data processing and data analysis and prediction. The collection of the tweets is often based on the public API of Twitter. It is a common practice to collect tweets by filtering according to specific queries, for example, using the name of the candidates in the case of elections [17]. Data processing includes all those data-curation techniques which aim to guarantee the credibility of the Twitter dataset. This is, for example, bot detection and spam removal [18]. Data analysis and prediction, the core of all these studies, can be simplified in four main approaches: volume analysis, sentiment analysis, network analysis and artificial intelligence (AI) including machine learning and natural language processing [13, 17–19].

Scholars used the number of mentions for a party of a candidate in order to forecast the result of the 2009 German parliament election [20]. While their technique has attracted many criticisms [21], their work was of inspiration to other researchers. Gaurav et al. [22] proposed a model, based on the number of times the name of a candidate is mentioned in tweets prior to elections, to predict the winner of three presidential elections held in Latin America (Venezuela, Paraguay, Ecuador) from February to April, 2013. Volumetric analyses have been also performed by Lui et al. [23] and Bermingham [24]. Ceron et al. [25] performed a sentiment analysis study on the tweets to check the popularity of political candidates in the Italian parliamentary election of 2011 and in the French presidential election of 2012. Caldarelli et al. [26] used the derivative of the volume to forecast the results of Italian elections. Scholars [27, 28] employed sentiment analysis to predict victory of Trump in the election of 2016 and Biden in that of 2020. Singh et al. also proposed a method [29] based on sentiment analyses and machine learning on historical data to predict the number of seats that contesting parties were likely to win in the Punjab election of 2017. It can be seen that sentiment analysis is widely used in election prediction [30]. Other works [31, 32] used analyses of social networks in order to identify the position of a party in the online community by measuring its centrality. The most supported parties are in general those with an higher centrality. Bovet et al. [17, 18] used a machine learning model based on in-house training set produced by hashtags from Twitter supporters to reproduce the polling trends leading to the 2016 US presidential election.

Despite the large amount of literature, the debate about whether Twitter or other social network outlets can be used to infer political opinions is still open. Online social networks are continuously filled by false, erroneous data through trolls, bots and misinformation campaigns to a level that distinguishing between what is genuine and what is not is in general difficult [18]. By virtue of this, the great challenge of algorithms and AI is to discover and interpret real data from "junk data" that could lead to accurate predictions of electoral or opinion trends. A crucial limitation of social network-based methods is also the mis-representation bias. While social networks solve the low response rates by "surveying" millions of users with non-intrusive methods, these large number of respondents might not represent, again, the demographics of the voting population. Thus, the opinions of Twitter users may not be representative of the entire population [33] and re-sampling methods need to be used, importing along them the same problems that plagued the traditional polls.

In this work we first investigate why the traditional polls fail to predict elections. We focus on the results of the primary presidential election in Argentina on August 2019 and the subsequent presidential election on October 2019, which represents a classic example of a massive failure of the trusted pollsters in predicting a polarized election electorate, which in this case, led also to large market collapses in the country, since investors largely bet on the pollster predictions. This study is possible thanks to the exclusive access to the raw data of longitudinal surveys[1] conducted by one of the most reliable pollsters in Argentina, Elypsis [34]. The analyzed data include the original

[1] The raw data of this pollster firm have been obtained by exclusive arrangement with the pollster responsible for conducting the polls of Elypsis (co-author Luciano Cohan) who has later founded his own company (Seido).

responses of subjects before performing the re-weighting for sampling bias and the subsequent results obtained after re-weighting. More importantly, the data includes a longitudinal study on the same 1900 respondents before and after the election which allows to precisely study the social-desirability bias when the same voter change the response after the result of the election is known. This represents a unique opportunity to discover why the traditional polls are failing, since pollsters do not normally share their raw data before re-weighting or sample-balancing [8, 9] and few surveys have been performed on the same respondents before and after an election.

We find that a poor demographic representation of the surveys combined with the inconsistency of opinion's respondents are the main reasons of the polls failure. We find a large mis-representation of the sample in the surveys as compared with the voting population, which in Argentina is approximately the general population since voting is obligatory and voter turnout is quite high at +80%. Even after re-weighting of the survey sample, this large sampling bias produces inaccurate results since important segments of society are highly underrepresented in the polls. Beyond this sampling problem, the main problem we find is a clear tendency for the respondents to not tell the truth about their preference for a candidate (Cristina Fernández) who was seen as controversial and the underdog in all polls and the news media. This social-desirability bias was the main culprit for the failure of the polls.

To overcome these problems, we propose an AI model to predict electorate trends using opinions extracted from social networks like Twitter. The machine learning mode consists of four phases: data collection, text and user processing, tweet classification with machine learning and opinion modeling. The model uncovers political and electoral trends without directly asking people what they think, but by trying to predict and interpret the enormous amount of data that people produce in online social networks [17, 18, 35]. Thus, this big-data analysis overcomes the low response rate problem of traditional surveys. By re-weighting the Twitter populations to the Census data, we match the distribution of the population' statistics and the statistics of the real demographic population (given by the Census Bureau) [33], thus minimizing the sampling bias of Twitter. Assuming that social network users freely express their opinions in social networks and since our methods are not interventionist, the data are, in principle, free of social-desirability bias. The real time data processing which underlies our AI algorithm allows us to detect sudden change of opinions in real time, and therefore different loyalty classes towards each candidate. We show that a cumulative longitudinal analysis tracking users over time performed on the loyalty classes to the candidates considerably improves previous forescasting results obtained in [17], which were based on instantaneous predictions. Instantaneous predictions, as well as pollsters predictions, are subject to high fluctuations which undermine the reliability of the prediction itself. Instead, here we show that taking into account the cumulative opinions of users over a long period of time produces a reliable predictor of people's opinion. These improvements allow us to obtain an accurate prediction on a difficult election, which dodged all pollsters in Argentina.

The contributions of the present study can be summarized into three aspects. Firstly, we find that a poor demographic representation combined with the inconsistency of opinion's respondents before and after the elections are the main reasons of the polls failure. Beyond this sampling problem, the main problem we find is a clear tendency for the respondents to not tell the truth about their preference. This social-desirability bias

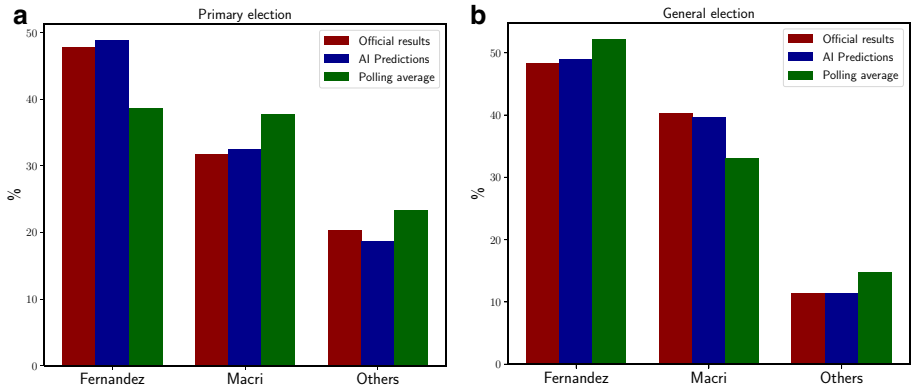

Fig. 1 Comparison between polling average (green), official results (red) and our prediction (blue) for both the (**a**) Primary (2019-08-11) and the (**b**) General election (2019-10-27) in Argentina 2019

was the main culprit for the failure of the polls. Secondly, we propose a novel machine learning model to predict electorate trends using opinions extracted from social network. By using machine learning and network homophily, we uncover electoral trends without directly asking people what they think. By re-weighting the Twitter populations to the Census data, we match the distribution of the population and the statistics of the real population thus minimizing the sampling bias of Twitter. Thirdly, we show that a cumulative longitudinal analysis tracking users over time performed on the loyalty classes to the candidates considerably improves previous methods. These improvements allow us to obtain an accurate prediction on a difficult election in Argentina. We validate the AI and network algorithms on the primary and general presidential election in Argentina 2019. Our results in this particular case show that social network with AI can capture the public opinion more precisely and more efficiently than traditional polls.

Why pollsters are failing to predict elections

The events leading up to the recent primary election in Argentina are a telling example of the failure of the polling industry [34, 36]. On the primary election day on August 11, 2019 (called PASO in Spanish: Primarias, Abiertas, Simultáneas y Obligatorias; in English: Open, Simultaneous and Obligatory Primaries), none of the pollsters in the country predicted the wide 16% margin of presidential candidate Alberto Fernández (AF) over the incumbent president Mauricio Macri (MM) (We clarify that primaries in Argentina are obligatory, happening for all political parties at the same time, and the two main parties presented only one candidate each, thus transforming the primaries into a de-facto presidential contest).

Figure 1 shows the comparison between the official results (in red), our prediction (in blue, Model 3 explained below) and the polling average, computed as the average of the top five most trusted pollsters in Argentina [34, 36], including Real Time Data, Management and Fit, Opinaia, Giacobbe and Elypsis (in green). Macri was clearly defeated by Fernández by +16%, a result captured by our predictions. While the average pollster predicted Fernández with a slight advantage in the primary, the pollsters estimated percentage for each candidates were, in general, really close reaching in some cases a difference of just one percentage point [37]. One of the most trusted pollsters, Elypsis,

who was the only one correctly predicting the previous presidential election in 2015, in particular predicted that Macri would win for one percentage point. This virtual tie predicted by the pollsters was largely considered to be a win for the incumbent candidate Macri, since he was supposed to gain all the votes of the supporters of the third party in the subsequent presidential election and, eventually, win the election in a runoff.

It is worth to stress at this point that Macri (right-leaning candidate) made the internationalization of the markets one of the main points of his campaign (favoring foreign investors and pro-business) while Fernández (left-leaning candidate) was instead supporting a pro-national market. As a result of the predictions supported by all Argentinian pollsters giving Macri as the winner, the bond market rose excessively in the days preceding the primary election. The subsequent defeat of Macri by 16 percentage points at the primaries led to a historic collapse of the MERVAL index by 40%, the bond market collapsed and some banks lost 1 billion dollars in the wager, overnight [38].

The failure of traditional polls is not associated with the impossibility of giving the exact right percentage for the candidates, but rather with the impossibility of predicting the enormous gap between them. The Argentina primaries elections are not the only example of pollsters' failure. Unpredictable results seem to be associated whenever one of the candidates is a controversial figure in the political scenario. In Argentina, the eventual vice-presidential candidate accompanying Alberto Fernández was previous Argentinian president Cristina Fernández de Kirchner (CFK), who had faced corruption allegations and judicial processes, and has often been vilified by traditional news media. Many Argentinians viewed her as a controversial and divisive figure in Argentinian politics. Other notorious examples of controversial candidates are Donald Trump in the US presidential election of 2016 and 2020 or Jair Bolsonaro in the Brazilian general election of 2018. In the case of the Argentina primaries, the pollster failure led to economic disruption of the country at a national/international level [38].

Below we analyze the raw data of one of the most reliable polls, Elypsis (trusted specially by the president Macri and international investors [34, 37, 38]) which, as all the pollsters, failed to predict the large gap between the two candidates for the primaries elections.

Figure 2a shows the age and gender distribution of the respondents to the survey conducted by Elypsis immediately before the PASO elections. Elypsis employs a combination of IVR of landline numbers complemented with online opt-in samples from Facebook. The vast majority of these online panels in Argentina, as well as in US, are made up of volunteers who were recruited online and who received some form of compensation for completing surveys, such as small amounts of money or frequent flyer miles. Figure 2b shows the number of respondents for Fernández (light blue), Macri (red) and Third Party (grey) grouped by age and gender. The Elypsis sample is peaked around the 50 years old group. This is strikingly different from the national population statistics obtained from the Argentinian Census Bureau shown in Fig. 2c.

Elypsis data does not have significant coverage among people younger than 30, even though it has been conducted in Facebook. It shows a heavier tail on the right for older groups, while the national population (Census Bureau) has a less pronounced peak around the group of 30 years old and a heavier tail on the left for younger groups. The largest sampled group surveyed by Elypsis are females between 51 and 65 years old who

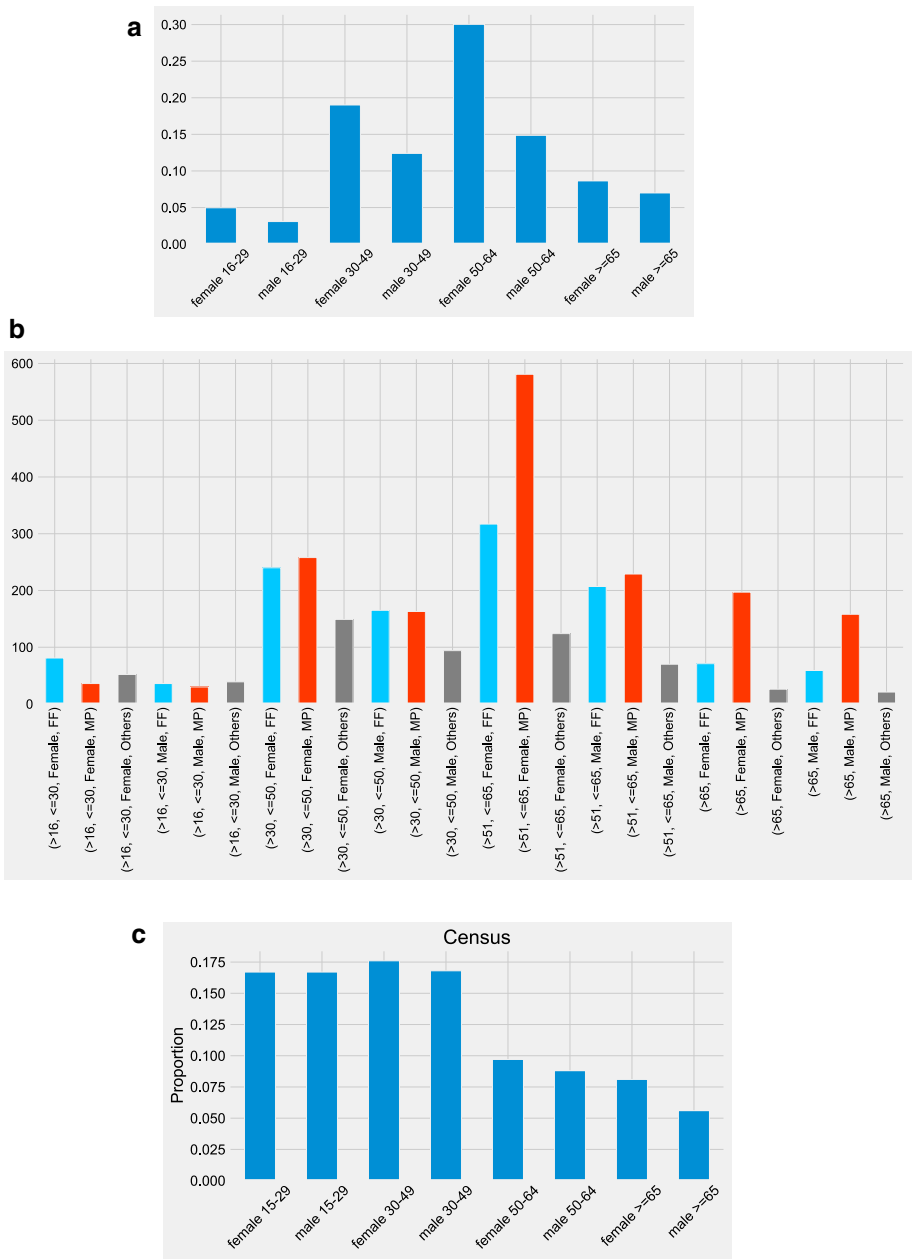

Fig. 2 **a** Elypsis polls demographic distribution of respondents to survey conducted before the primaries (PASO) by age and gender. **b** Elypsis polling results before PASO separated by age and gender (Macri=red, Fernández= blue, Others=grey). Results are highly biased to people older than 50 as compared with the Census distribution in panel **c**. **c** Argentina Census Bureau 2010 demographic distribution by age and gender used to compare the demographics of the polls and the Twitter sample

are overwhelming in favor of Macri. In fact, in all groups above 30 years old, Macri is the clear favorite in the Elypsis poll. On the contrary, Twitter represents better the younger generations. It is important to consider again that voting is obligatory in Argentina and citizens are allowed to vote as they turn 16 years old, and the turnout of the youngest is

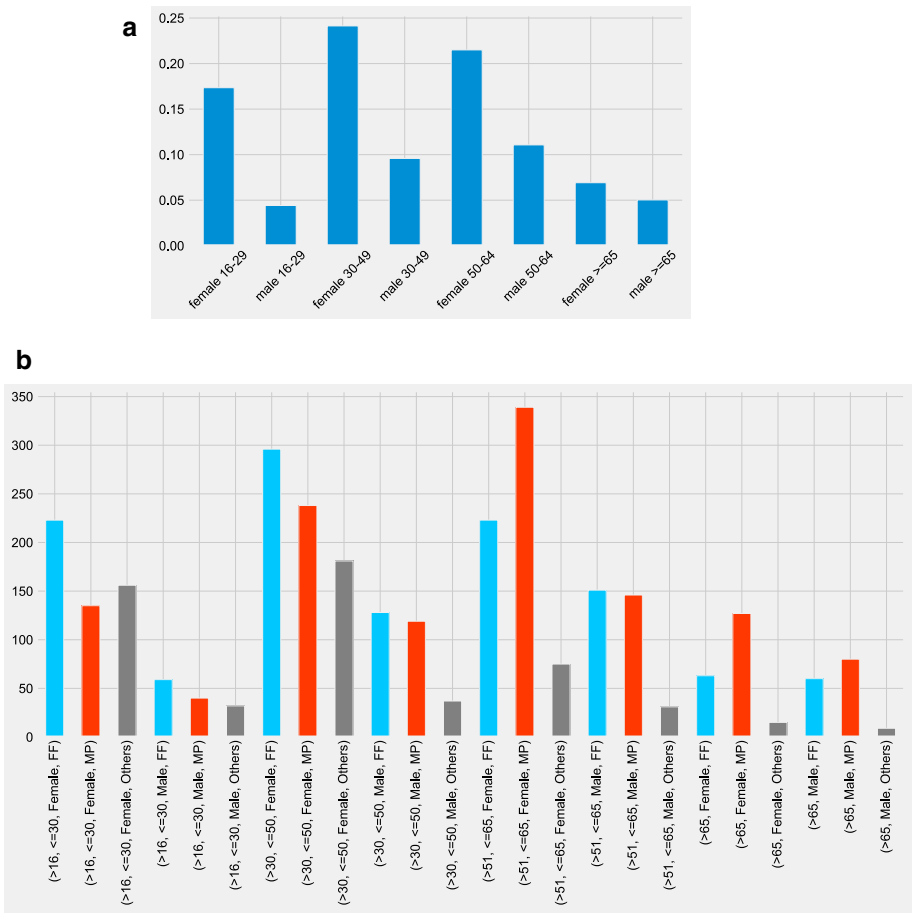

Fig. 3 **a** Elypsis demographics distribution of respondents to survey conducted after PASO by age and gender. **b** Elypsis polling results after PASO separated by age and gender (Macri = red, Fernándezvb= blue, Others = grey)

quite substantial, thus, any pollster who does not capture their preferences is, in practice, doomed to fail.

To deal with the mis-representation problem, pollsters adjust their raw data to population benchmark distributions given by the Census Bureaus [1, 2] by weighting the raw data (sample-balancing or raking [8, 9]). The poll sample is weighted so it matches the population on a set of relevant demographic or political variables, for instance, age, gender, location and other socio-economic variables, like education level or income. Studies of the effectiveness of various weighting schemes suggest that they reduce some (30 to 60%) of the error introduced by the biased sample, see [1]. However, when the raw data distribution is drastically under/over sampled as the Elypsis case, a small error in the most representative groups would propagate to produce inaccurate result.

As discussed above, the mis-representation is not the only problem which traditional polling methods face. Next, we analyze the longitudinal data taken on the same 1900 respondents by Elypsis before and after the elections to investigate the social-desirability bias. We start with Fig. 3a showing the Elypsis respondent distributions after PASO. Notice that this sample of respondents is different from the previous one,

and this is the reason why the age distribution changes with respect to the distribution before PASO. By comparing Fig. 2a before PASO with Fig. 3a after PASO we first notice a change in the voters distributions. Younger groups are better represented after the election when compared to Fig. 2a, although the data are still highly biased towards older generations. This implies that younger groups were, at least, more prone to answer the pollster after the election than before.

Surprisingly, the female group with ages between 30 and 50 years voted for Fernández as indicated after the PASO polls, while before the PASO they responded mainly in favor of Macri. The male group of the same age shows a similar behavior, even if less pronounced. Let us notice that, according to Fig. 2b, the groups of females/males between 30 and 50 years old are the most represented in the Census data and therefore have a higher impact on the final result. These results can only be explained by admitting that voters did not say the true.

This is further corroborated by this unique longitudinal panel, as seen in Table 1, revealing that people lied and hid their true voting intentions from the pollsters before the elections. More specifically, when comparing "Who are you going to vote in the PASO" with "Who did you vote in the PASO"—using the same sampling and post-stratification methodology than in the Pre-PASO survey—it is found that about 18% of the people did not disclose their true vote, and the hidden vote was not unbiased. Here, we refer to the real vote intention, i.e. the vote of each voter after PASO, as the true information.

- 91% of those who said "I will vote for Fernández" did so, but only 83% in the case of Macri, who lost 6% to AF.
- "Secondary candidate" voters were much more volatile, Only 56% of those who said that they were going to vote for (third candidate) Lavagna disclosed their true vote, and 54%, 53% and 59% in the case of other candidates Del Caño, Espert and Gomez Centurion respectively.
- Alberto Fernández got almost 19% of the votes of those who chose a secondary candidate in the Pre-PASO poll, and Mauricio Macri only 9%.

Table 1 Results obtained in the Elypsis surveys conducted before and after PASO. Results show a vote disclosure analysis of the respondents based on the two questions: "Who are you going to vote in the PASO" with "Who did you vote in the PASO"—using the same sampling and post-stratification methodology than in the Pre-PASO survey [46]

Who will you vote	Who did you vote			Was Pre PASO vote true	
	AF-CFK (%)	MM-MP (%)	Other (%)	Yes (%)	No (%)
AF-CFK	91	2	8	91	9
MM-MP	6	83	11	83	17
Lavagna	19	9	72	56	44
Del Cano	25	0	75	54	46
Espert	19	14	67	53	47
Gomez Centurion	10	8	83	69	31
Blank or Null	23	4	73	47	53
Unknown or others	53	11	36	*	*

Table 2 Results of the Elypsis surveys shown the hidden vote by demographics [46] of those who revealed to the pre-PASO pollster the candidate for which they voted in the election and those who did not revealed their true vote intention during the pre-PASO poll

	Revealed (%)	Not revealed (%)
Man	83	17
Woman	81	19
Between 16 and 30	67	33
Between 31 and 50	87	13
Between 51 and 65	90	10
More than 65	89	11
Full secondary	81	19
Incomplete secondary	81	19
Full or incomplete Univ.	86	14
Total	82	18

- Alberto Fernández received 46% of the votes of those who answered "Blank, Null or Unknown" before the PASO.

But, who hid - or not disclose - their real vote? We find no significant difference between men and women or between education levels but we see a clear pattern in age demographics. 33% of those between 16 and 30 years changed their response vs. their Pre PASO answer and only 13%, 10% and 14% on those between 31 and 50, 51 and 65 and more than 65, see Table 2.

What did those who did not disclose their vote think about the candidates? Where they "closeted Kirchnerists" (party of the previously mentioned politicians, AF and CFK) or did they bridge the gap between Macri-Fernández? "Regular" candidate images of Cristina Fernández, Macri and Alberto Fernández were much lower among those that did reveal their vote than among those who did not, see Table 3. Those who hid their votes look more nonpolarized, with a "Regular" image—No positive nor negative—of 21% on average, versus 6%/10% of those who revealed the vote. CFK's negative image is higher than MM (48% vs 38%) in "non-revealers" and the opposite hold in the revealers (43% vs 50%). 35% of the "non-revealers" did not have (or hid) their opinion of Alberto Fernández vs. 8% in the revealers.

This combined information shed light on PASO results and polls consensus miss. In the PASO, AF was able to catch votes from all the candidates, and seduce voters from

Table 3 CFK, AF and MM candidate Image in the Elypsis surveys as % of the total according to those respondents who revealed their intention of vote to the pollster and those who did not

Image % of the total	Revealed				Not revealed			
	Positive (%)	Negative (%)	Regular (%)	NS/NC (%)	Positive (%)	Negative (%)	Regular (%)	NS/NC (%)
CFK	45	43	6	5	20	48	22	11
MM	36	50	10	4	26	39	21	14
AF	45	39	7	8	15	28	28	35

within the gap, "moderate" voters who had a negative image of CFK and MM. He succeeded in standing himself as the "third candidate" bridging the gap, something that was not being fully captured by the polls, or that was decided at the last minute. This feature is most striking in young people, who may both have more "volatile" opinions and are less prone to reveal them on traditional polls.

This hidden-vote factor can explain by itself as much as 10% difference between "ex-ante" forecast and real results. Thus, standard poll methods failure may not have been related only to a bias in the sampling, but in the extraction of "True" information (i.e. the real vote intention) from surveyed people. Understanding why people lie to pollsters is not the topic of this work even if, according to the literature the reasons could be many and related to desirability-bias. On one hand, participants may typically rush through the surveys to obtain their rewards and don't respond thoughtfully [2]. On the other hand, social-desirability bias [11, 12], i.e. the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others [2, 12] is another reason for people to hide their preference for controversial candidates like CFK, which leads to biased results.

In view of how the above issues of low response rate, mis-representation and the social desirability bias/lies (which in the case of Elypsis biased more the younger representative) undermined the predictions of the Argentinian primary elections, we next search for suitable replacement techniques using sampling methods for the modern era of big-data science and AI. In this scenario, a good candidate to substitute traditional polls is the social network (Twitter in our study) which simultaneously solves the low response rate (millions of people express their political preferences in the microblogging platform) and the social desirability biases. This is because social network users do not answer to any question, but freely express their ideas in a social medium platform. However, one may argue that Twitter is generally biased towards young people thus providing a biased sample. Thus, proper re-weighting of the data is needed, although the effects of re-weighting are found to be less pronounced than in the polls of Elypsis. Below we introduce an AI model that builds up on previous work in [17] combining machine learning, network theory and big-data analytic techniques, that is able to overcome the problems presented so far and that correctly predicted the outcome of the 2019 Argentina primary and general elections.

Method

The algorithm we propose improves upon previous work from [17] and consists of four phases (see Fig. 4): data collection, text and user processing, tweet classification with machine learning and opinion modeling. While the first two phases are of standard

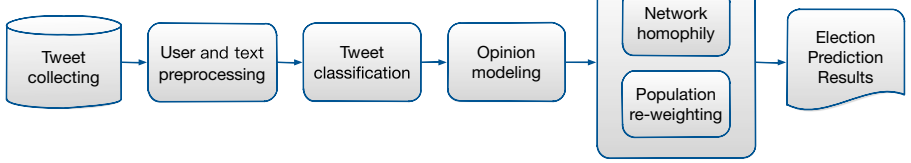

Fig. 4 The flow of the election prediction algorithm relying on machine learning and network theoretical tools for opinion modeling

practice in the literature, tweet classification by means of machine learning only recently took place [17, 18], given the impossibility to classify by hand millions or even billions of datapoints. Opinion modeling, the core of our election prediction model, is an attempt to instantly capture people's opinion through time by means of a social network. To improve upon [17], we consider the cumulative opinion of people and define three prediction models based on different assumptions on the loyalty classes of users to candidates, homophily measures and re-weighting scenarios of the raw data. Below we explain each phase, highlighting the steps that make our full-fledged AI predictor a good candidate substitute for the traditional pollster methods.

Data collection

By means of the Twitter public APIs, we continuously collect tweets until the election day (from March 1, 2019 until October 27, 2019), filtered according to the following queries (corresponding to the candidates' name and handlers of the 2019 Argentina primary election): *Alberto AND Fernández*, *alferdez*, *CFK*, *CFKArgentina*, *Kirchner*, *mauriciomacri*, *Macri*, *Pichetto*, *MiguelPichetto*, *Lavagna*. Only tweets in Spanish were selected. Figure 5a shows the daily volume of tweets collected (brown line) while Fig. 5b shows the daily number of users (green line). In blue we report the daily number of tweets/users which are classified, i.e. they posted at least one classified tweet. Users are

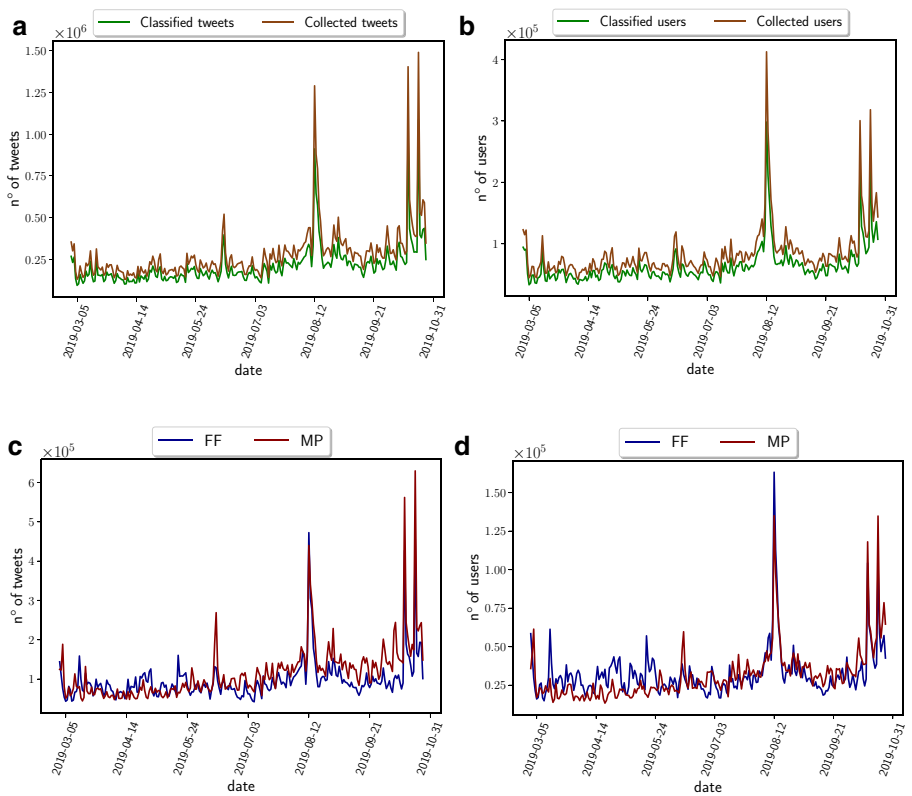

Fig. 5 **a** Daily volume of collected (brown line) and classified (green line) tweets. **b** Daily volume of collected (brown line) and classified (green line) users. **c** Daily tweets supporting the FF and MP formula. **d** Daily users supporting the FF and MP formula

classified with machine learning (explained below) as supporters of Macri (Fig. 5d, red line) if the majority of their daily tweets are classified in favor of Macri (Fig. 5c, red line) or as supporters of Fernández in the other way around (blue line in Fig. 5d and c). Hereafter we use FF to indicates the Fernández-Fernández formula and with MP we refers to the Macri-Pichetto formula (the outgoing president/vice-president candidate).

The activity of tweets/users shows a peak on August 11, 2019, i.e. the day of the primary election. In the period from March to October, we collected a daily average of 282,811 tweets posted by a daily average of 84,062 unique users. We daily classified 75% of these tweets and ∼ 76% of the users (see Additional file 1: Table S1 and Table S2). In total, by the end of October we collected around 110 million tweets broadcasted by 6.3 million users. This large amount of tweets collected has no precedent and is relevant in the light of considering that Argentina is one of the most tweeting per capita countries in the world.

User and text processing

Below we explain the tasks that need to be applied to the raw data before any analysis is performed.

Bot detection

The identification of software that automatically injects information in the Twitter' system is of fundamental importance to discern between "fake" (bots) and "genuine" users, the latter representing the real voters.

According to [18] a good strategy is to extract the name of the Twitter client used to post each tweet from their source field and kept only tweets posted from an official Twitter client. Third-party clients represents a variety of applications, form applications mainly used by professionals for automating some tasks (e.g. dlvrit.com) to manually programmed bots. This simple method allows to identify tweets that have not been automated from those automated by bots, and scales very easily to large datasets contrary to more sophisticated methods.

Figure 6a and b show the daily number of tweets posted by bots and the daily volume of bots, respectively. Figure 6c and d show the daily volume of classified tweets/bots. The daily average of bots between March and October 2019 is 732 with an overall daily activity (in average) of 2243 tweets. The daily classified tweets are 1617 while the daily classified bots are 560 bots. As for "genuine" users, a bot is classified if it share at least 1 classified tweet. In the entire dataset we found around 20,000 bots which posted 538,350 tweets. Let us notice that even though we classified the bots, they are not used for the final prediction since they do not corresponds to real voters.

Text standardization

Stop word removal and word tokenization are of common practice in Data mining and Natural language processing (NLP) techniques [39, 40]. For example, we keep the URLs as tokens since they usually point to resources determining the opinion of the tweet, through replacing all URLs by the token "URL".

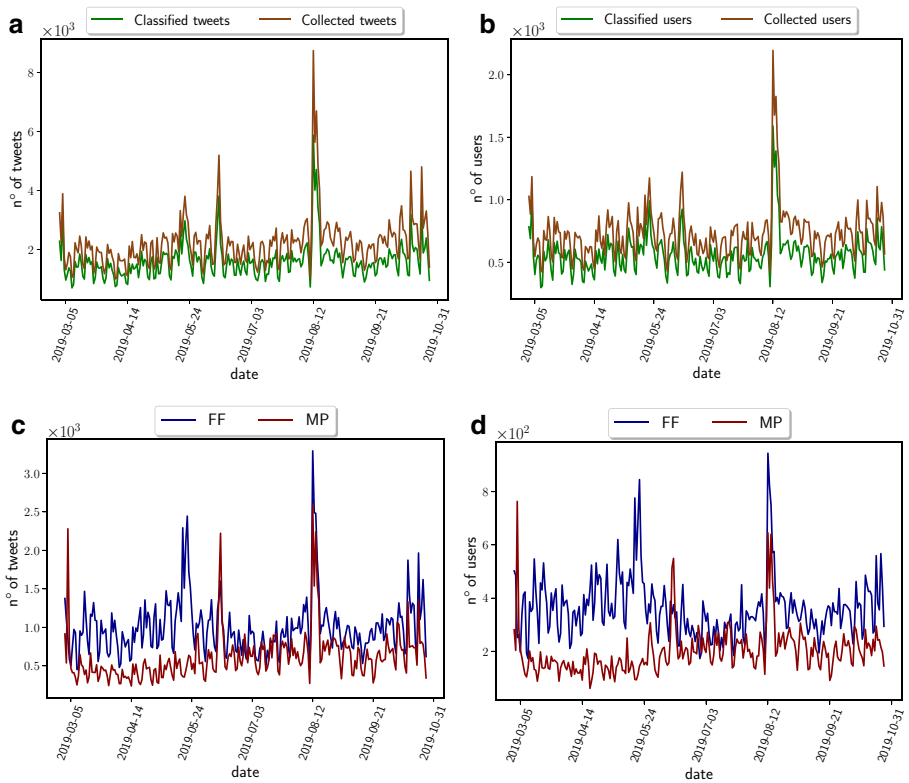

Fig. 6 Bots analysis: we detect bots with methods from [18] and classify them according to bots favorable to FF or MP. **a** Daily volume of collected (brown line) and classified (green line) tweets from bots. **b** Daily volume of collected (brown line) and classified (green line) bots. **c** Daily tweets from bots supporting the FF or MP formula. **d** Daily bots supporting the FF or MP formula

Tweet classification

To build the training set we analyze the hashtags in Twitter. Users continuously labels their tweet with hashtags, which are acronyms able to directly transmit the user feeling/opinion toward a topic. We manually labeled the top hashtags used in the dataset in 2019 August (see Additional file 1: Table S3, also we report hashtags after August in Additional file 1: Table S4). They are classified either as pro Macri, pro Fernández or pro Third party candidate, depending on who they support (with Third party we refer to the supporters of Lavagna, Espert and other secondary candidates). We also consider negative hashtags for one candidate as pro for the other candidate, when this is obvious. For instance, #Macriescaos as pro-Fernández.

Hashtag co-occurrence network

In order to check the quality of the classification of the classified hashtags we build the hashtag co-occurrence network $H(V, E)$ and statistically validate its edges [17, 41]. In the co-occurrence network the set of vertices $v \in V$ represents hashtags, and an edge e_{ij} is drawn between v_i and v_j if they appear together in a tweet. We test the statistical significance of each edge e_{ij} by computing the probability p_{ij} (p-value of the null hypothesis) to observe the corresponding number of co-occurrences by chance only knowing

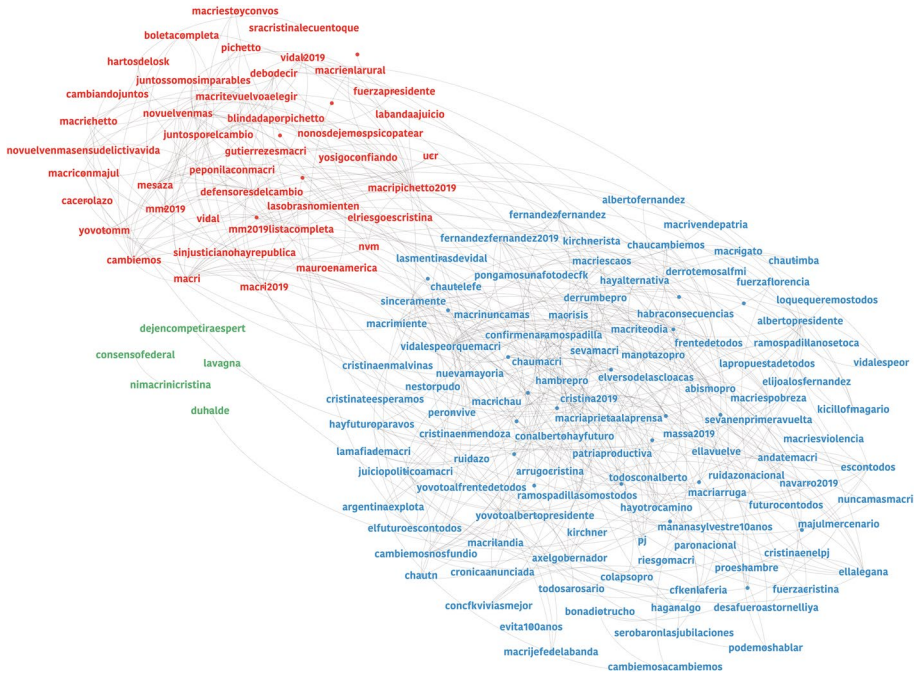

Fig. 7 Hashtag co-occurrence network from March to August 2019. We display in blue the hashtags in favor of Fernández-Fernández, in red the hashtags in favor of Macri-Pichetto and in green those in favor of the Third party

the number of occurrences c_i and c_j of the vertices v_i and v_j, and the total number of tweets N. Fig. 7 shows the validated network. We only keep those edges with a p-value $p < 10^{-7}$. The blue community contains the hashtags in favor of Fernández, the red community those in favor of Macri and the green one (a very small group) are those in favor of the Third candidate. A look at the typologies of hashtags reveals the first differences in the supporters. Those in favor of Cristina Kirchner are much more passionate than the follower of Macri. For example, Kirchner's type of hashtags are #FuerzaCristina, #Nestorvuelve, #Nestorpudo or they are very negative to Macri as #NuncamasMacri. On the other hand, Macri's group is smaller and less passionate with hashtags like #Cambiemos or #MM2019 (see Fig. 8), while support for the third candidate has not taken traction and its electoral base on Twitter is very small.

In principle, counting the users and tweets according to the hashtags they use would predict the victory of Fernández over Macri. However this conclusion would be based only on $\sim$ 10,000 users (those expressing their opinion through hashtags). In order to get the opinion of all the users we train a machine learning model that classifies each tweet as FF, MM or Third party (In what follows we also refer to the formulas FF for Fernández-Fernández and MP for Macri-Pichetto, the final formulas in the presidential contest). We use the previous set of hashtags expressing opinion to build a training set of labeled tweets, which are used in turn to train a machine learning classifier. We use all the tweets (before August) which contain at least one of the classified hashtags to train the model. In the case of more than one hashtag for a tweet, we consider it only if all the hashtags are in favor of the same candidate.

Fig. 8 Hashtag clouds. The dimension of the words is proportional to their frequency in the dataset. **a** The blue hashtags are the most frequently used in the tweets in favor of Fernández. **b** The red hashtags are those in favor of Macri

The use of hashtags that explicitly express an opinion in a tweet represents a "cost" in terms of self-exposition by Twitter users [42] and therefore allows one to select tweets that clearly state support or opposition to the candidates. The training set consists of 228,133 tweets, i.e. the 0.33% of the total amount of collected tweets and the ~ 90% of the hand-classified tweets (253,482 tweets). In order to find the best classifier we used five different classification models, the Logistic Regression (LR) with L_2 regularization, the Support Vector Machine (SVM), the Naive Bayes (NB), the Random Forest (RF) and the Decision Tree (DT). All these models are validated on the remaining 10% of the classified tweets (25,349). Table 4 shows the results for the models. The Logistic Regression performs better than the other models with an average group accuracy equals to 83%. Also recall and F1-score are equal to 83%. Support Vector Machine is the second best classifier, with an average accuracy of 81%, followed by the Naive Bayes with and average accuracy of 79.5%, the Random Forest and the Decision Tree.

Table 4 Performance of the AI classification models: Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) and Decision Tree (DT)

Model	Precision (FF)	Recall (FF)	F1 (FF)	Precision (MP)	Recall (MP)	F1 (MP)
LR	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**
SVM	0.81	0.81	0.81	0.81	0.80	0.81
NB	0.79	0.80	0.80	0.80	0.79	0.80
RF	0.74	0.80	0.77	0.79	0.72	0.75
DT	0.76	0.76	0.76	0.76	0.76	0.76

The Logistic Regression performs better than the other models with an average group accuracy equals to 83%

We recall that the logistic regression assigns to each tweet a probability p of belonging to a class. In our case such probability goes to one if the tweet supports Macri while it goes zero if it supports Fernández. As it is shown in Fig. 9 the distribution of p contains two peaks, one on the left and one on the right, divided by a plateau. This is an encouraging result, since it proves the efficacy of the model to discern between the two classes. We classify a tweet in favor of Macri if $p \geq 0.66$, in favor of Fernández if $p \leq 0.33$, otherwise it is unclassified. Tweets with a value of p in the plateau are instead unclassified, meaning that the tweet does not contain sufficient information to be classified in either camp. According to this rule, in average we classify 211,229 genuine tweets and 1617 tweets from bots per day (see Additional file 1: Tables S1, S2).

Opinion modeling

We can infer users' opinion from the majority of the tweets they post. Let $n_{t,F}$ be the number of tweets posted by a given user at time t in favor of Fernández and let $n_{t,M}$ be those supporting Macri. We define an instantaneous opinion over a window of length w and a cumulative average opinion as follow. In the first case, a user is classified as a supporter of Fernández (at a given day $t = d$) when

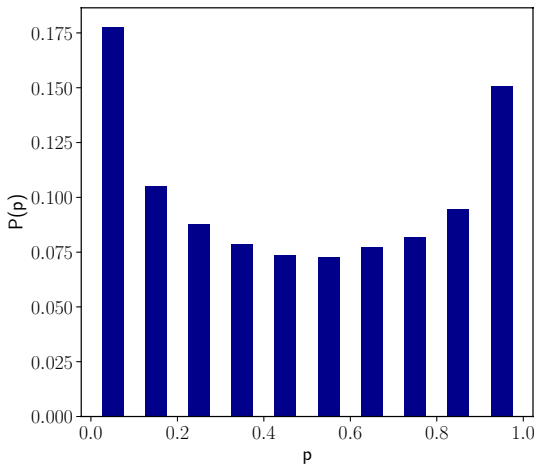

Fig. 9 Probability distribution of p = probability of voting for Macri ($p = 0$ corresponds to Fernández) obtained by the Logistic Regression model

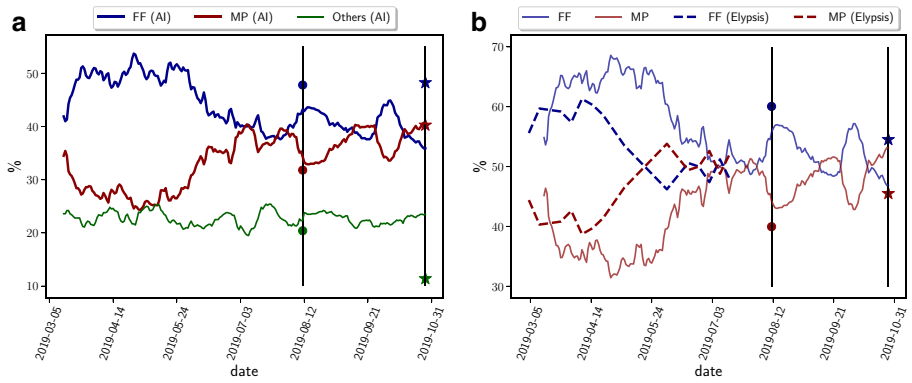

Fig. 10 **a** Results from *Model 0*. Instantaneous results of the AI *Model 0* obtained in a moving window of w=14 days. Vertical lines are (from the left to the right) the day of the primaries and general elections respectively. The circles represents the official results for the primaries while the stars are those for the general election. **b** Results from **a** compared with polls from Elypsis. Thick continuos lines represent AI results from *Model 0* (instantaneous) while dashed line represent the Elypsis predictions. The instastaneous model does not predict the results of the elections. Instead, we consider the cumulative predictions in Figs. 11 and 12

$$\sum_{t=d-w+1}^{d} n_{F,t} > \sum_{t=d-w+1}^{d} n_{M,t} \tag{1}$$

i.e if the majority of the tweets posted in the last w days were in favor of Fernández. The user is classified as a supporter of Macri if

$$\sum_{t=d-w+1}^{d} n_{F,t} < \sum_{t=d-w+1}^{d} n_{M,t} \tag{2}$$

If none of the previous conditions is met, i.e. if

$$\sum_{t=d-w+1}^{d} n_{F,t} = \sum_{t=d-w+1}^{d} n_{M,t} \tag{3}$$

then the user is classified as undecided. Let us notice that when w goes to one we have the 'most' instantaneous prediction, that is the prediction based on what people think in the last day. This instantaneous prediction model was used in the previous work [17] to match the results of the AI model to the aggregate of polls from the New York Times in the 2016 US election with excellent results. However, this predictor did not match the results of the electoral college, which required stratification by states. Thus, we further develop the AI model of [17] to add other predictors beyond the instantaneous measures.

Traditional poll data collection is an instantaneous prediction with a value of w that can go from few days up to few weeks, which is the time of collection of the poll data and this corresponds roughly to our instantaneous measurement above. However, the fact that we are able to track the same user over long period of time in Twitter allows us to extend the window of observation as far as we want to then define a new measure that we call the 'cumulative opinion'. The cumulative opinion in our model is defined by

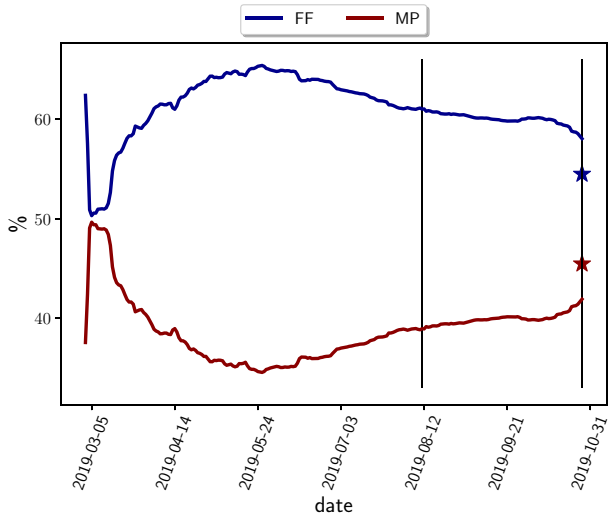

Fig. 11 Results of the cumulative *Model 0* defined without considering the loyalty classes

extending w to the initial date of collection for every time d of observation, i.e., $w = d$. Thus the cumulative opinion considers the opinion of a user based on all the tweets he/she posted from time $t = 0$ up to the observation time d. That is, our prediction is longitudinal as we are able to follow the opinion of the same user over the entire period of observation of several months. In terms of traditional poll methods, a cumulative opinion would be obtained in a group panel collecting for each respondent in the sample and for each day starting from $t = 0$ her/his preference towards a candidate. This possibility, which would require an unimaginable amount of effort and time for traditional poll methods, it is quite straightforward when it comes to social networks and big-data analyses.

We start by investigating the instantaneous response of the users in a fixed window of time. Figure 10a shows the Twitter supporters dynamics over time obtained with a window average, $w = 14$ days. Users are classified as MP (Macri-Pichetto, in red), FF (Fernández-Fernández, in blue) or Others (in green). Figure 10b shows the supporters dynamic (thick lines) compared with Elypsis prediction from their pools (thin dashed lines) without considering the undecided users in the normalization. In the same plots we also report the official results for both primaries and general elections. The comparison between the two figures stands out as an approximate correlation between the Elypsis and our AI results for each candidate valid for these instantaneous measures. However, in the comparison among candidates, predictions may sometimes differs, as for example, right before the beginning of August, Elypsis gave as favorite MP while the AI instantaneous prediction was in favor of FF. Overall, as for the pollsters results, window average analyses are representative of the instantaneous sentiment of the people. As we see from the figures, instantaneous opinions are affected by considerable fluctuations [17] which make the prediction not reliable. In Additional file 1: Fig. S1, we compare the average window opinion with other pollsters (Real Time Data, Management & Fit, Opinaia, Giacobbe and Elypsis). An interpolation (thin lines) shows similar trends as the AI-model window average, stressing that the conclusions made so far are more general than

the single comparison with Elypsis. In fact in [17] we have shown that the instantaneous predictions of the AI model follows quite closely the aggregation of polls obtained from the New York Times, 'The Upshot', yet, it does not reproduce the results of the general election in the electoral college which further requires a segmentation by states where proper prediction of rural and non-rural areas becomes the key. Considering the cumulative opinion, not the instantaneous one, of each user is crucial to correctly predict the elections.

Thus, we next study the opinion of each user by considering the cumulative number of tweets over the entire period of observation to classify the voter's intention (*Model 0*). This cumulative approach takes into consideration all the tweets together for each user since the first time they enter in the dataset and bases the voter intention on all of them. This cumulative approach can only be done with Twitter and not with traditional polls, except for short times and particular cases as done by Elypsis before and after PASO.

Figure 11 shows the prediction of the model using the cumulative opinion of the users from March 1 until a few days before the general elections. We can see that this approach captures the election results well, and, in particular, the huge gap between the candidates, both for the primary election and the general election (vertical lines from the left to the right). While a low precision is of secondary importance when the difference between the opponents is high, it plays a central role when they have a close share of supporters. As an extreme example, in an almost perfect balanced situation, the change of mind of just few people may flip the final outcome. If on the one hand a cumulative approach do reduce the fluctuations in the signal, it is also less sensitive to sudden change of opinion. A person can support a candidate until few days before the elections, and then change her/his mind because of some particular facts. This and other possibilities can be taken into account only by a model based on cumulative analyses, able to capture the degree of loyalty of people towards the candidates over time. Differently from the traditional surveys, the real-time data processing that underlies our AI algorithm gives the possibility to take into consideration this scenario. To understand how different re-weighting scenarios affect the results, below we introduce different loyalty classes of users towards the candidates and then we define several models matching the criteria previously discussed. These loyalty classes can be defined when we consider the cumulative opinion in a longitudinal study and cannot be investigated by traditional polls.

Loyalty classes

We define five classes of loyalty for users who support candidate c. Here we consider the FF supporters, but the definitions below similarly applied to the other candidates.

- Ultra Loyal (UL): users who always tweet only for the same candidate, for example

$$\sum_{t=T_0}^{T} \left(\frac{n_{F,t}}{n_{F,t} + n_{M,t} + n_{T,t}}\right) = 1, \tag{4}$$

  where with $n_{c,t}$ we indicate the number of tweets that a given user post in favor of c, with $c \in \{\text{M (Macri), F (Fernández), U (Unclassified)}\}$.

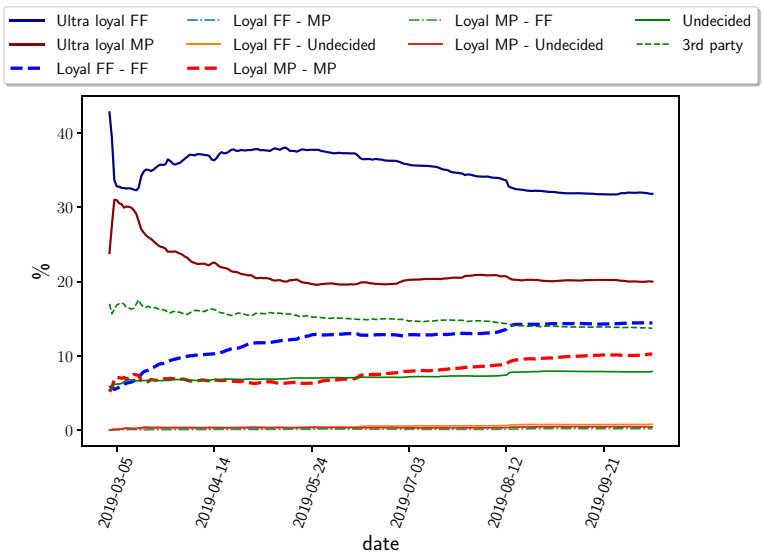

Fig. 12 Cumulative users' opinion trends for each loyalty class over time

Differently from the ultra loyal, which continuously post in favor of a candidate, the other classes take into consideration a possible change of opinion of a user. In order to detect sudden twist of opinions we focus on the classifications of the last k tweets posted by the users. We define:

- Loyal FF → FF: a user which is FF since the majority of tweet are for FF, but she/he also supported FF in the last k tweets, mathematically speaking

$$\sum_{t=T_0}^{d} n_{F,t} > \sum_{t=T_0}^{d} n_{M,t} \; AND \sum_{i=N-k+1}^{k} n_{F,i} > \sum_{i=N-k+1}^{k} n_{M,i} + n_{U,i} \tag{5}$$

  N is the total number of tweets posted by the user and $n_{c,i}$ is 1 if the i-th tweet of user is classified as supporting c, otherwise 0.
- Loyal FF → MP: users that are FF by the total cumulative count but they have tweeted for MP in the recent k tweets. namely in formula:

$$\sum_{t=T_0}^{d} n_{F,t} > \sum_{t=T_0}^{d} n_{M,t} \; AND \sum_{i=N-k+1}^{k} n_{M,i} > \sum_{i=N-k+1}^{k} n_{F,i} + n_{U,i} \tag{6}$$

- Loyal FF → TP: users supporting the third party in the last k tweets, i.e.

$$\sum_{t=T_0}^{d} n_{F,t} > \sum_{t=T_0}^{d} n_{M,t} \; AND \sum_{i=N-k+1}^{k} n_{U,i} > \sum_{i=N-k+1}^{k} n_{F,i} + n_{M,i} \tag{7}$$

- Loyal FF → Undecided: all other individuals classified as FF but not included above.

Let us remind that unclassified refers to all those users who do not have any classified tweet. Fig. 12 shows the cumulative prediction for each class, with T_0= March 1, 2019

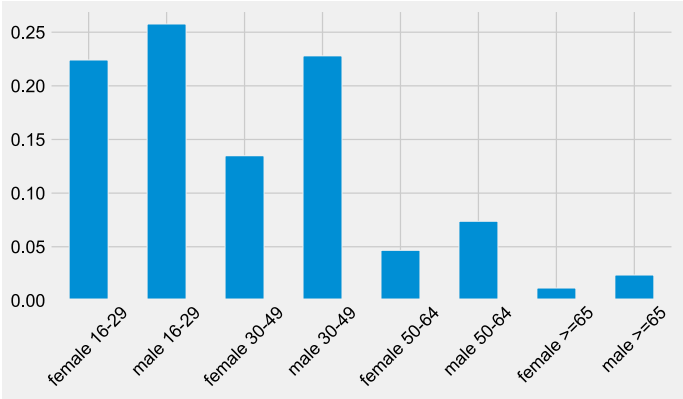

Fig. 13 Distribution of Twitter populations by age and gender during 2019 Argentina election

and $k = 10$. The Ultra Loyal class for Fernández (FF) represents $\sim$ 33% of the populations while only $\sim$20% of the populations is Ultra Loyal towards Macri (MP). Loyal MP→MP and loyal FF→FF represents between the 8% and the 13% of the studied Twitter population. The percentage of the undecided is around 8% and the third party percentage. The other classes are close to 1 or 2%. In the next section we use these classes in order to define a better election predictor than Model 0.

AI models based on loyalty classes

The loyalty classes introduced so far are one of the main differences with the Twitter based studies from [17]. Here, we use the machine learning classifier (logistic regression here) to define the loyalty of a user and not to make predictions. We then make the predictions by grouping supporters as follows:

- Fernández supporters: all those users who are Ultra loyal FF, Loyal FF→FF, Loyal FF→MP, Loyal FF → Undecided.
- Macri supporters: all those users who are ultra loyal MP, Loyal MP→MP, Loyal MP→ FF, Loyal MP→Undecided.

In each group we put those users we are almost sure who they support because of their activity over time. However, as we saw in the previous section, undecided may play a central role in a scenario where few percentage points can flip the final result. Furthermore, understanding unclassified users (i.e. those users which do no not have any classified tweet) will also improve the final statistic. In order to take into account all the reasonable scenarios of results, we define three different models (starting from the classification in Fernández and Macri above) and validate them against the final results of the election. Table 5 resumes the details of each model.

Model 1: All the users belonging to one of the following classes are grouped in the Third Party: Undecided→MP, Undecided→FF, Undecided→Undecided and Unclassified.

Model 2: Instead of simply grouping the undecided in a third party, we use network homophily to infer their political orientation. A user is classified as MP(Undecided) if the majority of her/his neighbors in the undirected retweet network supports

Table 5 The definition of three AI forecasting models based on the loyalty classes and network theoretical tools of homophily according opinion modeling

MODEL 1	**Supporters (classes of users)**
FF	Ultra loyal MP, loyal MP→MP, loyal MP→FF and loyal MP→Undecided
MP	Ultra loyal FF, loyal FF→FF, loyal FF→MP and loyal FF→ Undecided
Third party	Undecided→MP, Undecided→FF, Undecided→Undecided, Unclassified
MODEL 2	Supporters (classes of users)
FF	Ultra loyal FF, loyal FF→FF, loyal FF→MP, loyal FF→Undecided and FF(undecided)
MP	Ultra loyal FF, loyal FF→FF, loyal FF→MP, loyal FF → Undecided and MP(undecided)
Third party	Undecided(undecided), Unclassified
MODEL 3	Re-weighting the MODEL 2 from the Twitter populations to the Census data

Macri. The same definitions applied for the other cases. In this model, FF(Undecided) are considered supporters of Fernández and MP(Undecided) supporters of Macri. Undecided(Undecided). In this model the Unclassified belong to the Third Party (or Others). We remind the reader that the Unclassified users are those users who always tweeted unclassified tweets (those with $0.33 < p < 0.66$). Undecided users are those who satisfies Eq. (3).

Model 3: We are also analyzing the users' profiles while collecting tweets, including the head portrait and the location information. Firstly, users outsides Argentina are removed from the dataset. Face analyzing tools developed by a state-of-the-art facial recognition algorithms have been validated and applied in the context of profile pictures [43, 44]. We obtain more than 400 thousand users' age and gender information, see the population distribution of Twitter users from Fig 13. In this model, we adjust the results of Model 2 for each candidate by re-weighting the Twitter population to the Census data by

$$\sum_{gender,age} \frac{N_{u_c}(gender, age)}{N_u(gender, age)} \frac{N_{\text{Census}}(gender, age)}{N_{\text{Census}}} \tag{8}$$

$N_{u_c}(gender, age)$ user, whose gender equals *gender* and age equals *age*, support candidate *c*. From the Census data, $N_{Census}(gender, age)$ people belong to the category of *gender* and *age*. ($c = \{\text{FF, MP, Third party}\}$ $gender \in \{\text{female, male}\}$, age $\in \{[16, 30], [31, 50], [51, 65], [65, \infty]\}$).

In the next section we employ these three models and compare the performances of models on the 2019 Argentina election.

AI-based forecast for the Argentinian election

The models introduced so far allow us to define the daily supporters of each candidate according to their Twitter activity. Indeed supporters are defined not simply according to the classification of the majority of their tweets, but on the basis of the loyalty classes they belong to. Similarly to the simple tweet classification, we can define for each model an instantaneous (window average) and a cumulative (average) opinion.

Figure 10 shows that an instantaneous indicator provides an approximate fitting to the results of polls. We have already used this indicator, in our previous study of the 2016 US

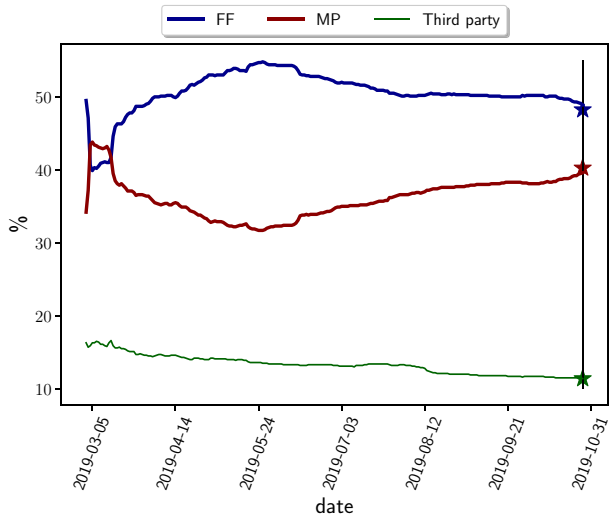

Fig. 14 Cumulative *Model 3* prediction of the AI model and comparison with primary on August 11, 2019 and general election results on October 27, 2019. *Model 3* is the best fit to the real data

presidential election, to precisely fit the New York Times Aggregator of Polls at The Upshot [17, 45]. This aggregator unifies a thousands polls and weights them with proprietary information to produce a weighted average of all the most trustable pollster in USA. While this analysis is interesting and give the opportunity to predict instantaneous changes in electoral opinion, this indicator does not provide the electorate opinion as a whole and it is not the most important predictor of the election outcome. It is not the greatest information that can be extracted from social networks, either, and indeed, it failed to predict the US 2016 election and the present Argentina 2019. The estimator that predicts better the election is provided when we consider the cumulative number of users from the beginning of measurements, and not just the behavior of the users in a small window of observation.

For this reason here we directly focus on the cumulative prediction for the models introduced in the previous section. Table 6 reports the prediction of each model right before the day of the general election on October 27, 2019. The official results saw the victory of Fernández with 48.24%. Macri scored 40.28% and the Third Party with 19.48%. The average predictions obtained by averaging the results of the three models are consistent (inside the standard error) with the official outcome. Indeed, we obtain $(48.0 \pm 1.5)\%$ for FF, $(35.4 \pm 3.1)\%$ for MP and $(16.6 \pm 4.1)\%$ for the Third Party (TP). This in an interesting result which highlights the importance of considering loyalty classes for political elections. In order to establish the best among the three models we compute the mean absolute error between each model' prediction and the final results.

Let $Y = \{y_c\}$ with $c \in \{\mathrm{FF}, \mathrm{MP}, \mathrm{TP}\}$ be the prediction of one model and let $X = \{x_c\}$ with $c \in \{\mathrm{FF}, \mathrm{MP}, \mathrm{TP}\}$ be the official results. We define the MAE_i (mean absolute error) for model i as

$$\frac{\sum_{i \in c}(|x_i - y_i|)}{3} \tag{9}$$

Table 6 Prediction results for the general election on October 27 of AI Models 1, 2, and 3 as defined in the text

MODEL	FF (%)	MP (%)	Third party (%)	MAE (%)
1	45.9	32.5	21.6	6.71
2	49.1	34.0	16.8	4.21
3	**48.9**	**39.6**	**11.5**	**0.53**

Table 6 shows the MAE for each model. *Model 3*, based on the homophily detection and population re-weighting is the best predictor with a mean absolute error of 0.53. This model predicted 48.9% for FF (an overestimation of 0.66 points if compared to the official result), 39.6% for MP (an underestimation of 0.68 points) and 11.6% for the Third Party. As a matter of fact, the AI model 3 is capable of predicting the Argentinian general elections, by giving a percentage of electors for each candidate close to the official one and outperforming traditional poll methods, as is shown in Fig. 1b and in Fig. 14.

Maybe the most important result is the performance of our algorithm before the PASO, where all the pollsters failed to predict the +16% points difference between the two candidates (by strongly underestimating their gap). *Model 3* predicts a difference of almost 18% points in favor of FF, close to the official result.

While the PASO primary results could be considered of secondary importance, they play a central role in the Argentina political campaign and they are the most difficult to guess because it's the first time the citizens officially express their opinion on the election. Figure 1 shows how traditional pollsters modified their prediction after the primary elections, somehow fitting them with the PASO results. How they modified their predictions is still not clear and some of the pollsters (Elypsis for example) did not release any prediction after the PASO.

A study of the hashtags and queries of the followers of the FF formula indicates that the vast majority of the people focused more on the poor economic situation in which the country was at the time of the election, instead of the judicial cases of corruption that affected the FF candidates. Most of the hashtags reflects sentiment of hunger, chaos, crisis and despair. On the other hand, the expression of the followers of Macri-Pichetto is reflected in hashtags that give strength to the president, but they do not reflect a feeling for the economic and political situation, but more a moral support, perhaps of resignation. The followers of Macri do not express too much their concerns about judicial cases of corruption either (Fig. 8).

Finally, let us notice that the cumulative average depends on the initial time T_0. This value determines the initial fluctuations of the cumulative average, which generally stabilize into a value that it is difficult to change unless a big swing in opinion of the electorate suddenly appears. To investigate this effect, we have recalculated the cumulative average by changing the origin of measurement T_0. As we see from Additional file 1: Fig. S2, the predictions for the general elections cluster around the same value. We use this fluctuations to compute the error associate to our final predictions. We define the error as the standard deviation over the results of different realizations with $t < T_0$. Regarding *Model 3*, the estimated average error is 0.53%. This result strengthens the goodness of our prediction, which are consistent, inside the error bars, with the final results.

Conclusion

One of the fundamental tools of artificial intelligence in social networks is that it captures changes in people's opinions without any intervention and for an extended time. Then AI can capture the sentiment of the millions of users who constantly express themselves on the internet and change or maintain their positions. AI can also filter this information from manipulators and bots and can reduce it to its essence, by overcoming the problems traditional pollsters face: low response rate, social desirability biases and the mis-representation of the population.

The results of our analyses show that AI applied to big data can be used to successfully understand people' opinions over time. We show that the possibility of following the opinion of the same people through time, and therefore the chance of defining loyalty classes is an important step in order to make good predictions. AI allows, both, to get the percentage of supporters towards a candidate and reveals what is behind these numbers, giving an idea of people sentiments. This is of particular importance when one of the candidates is a controversial politician and can generate different feelings leading to strong polarization and biased responses to pollsters, which are not trusted anymore by the great majority of people.

This study has inevitable limitations, especially for the prediction model, which might be interesting directions in the future work. For example, how influencers in Twitter affect users' opinion is not fully discussed and we will consider influencers and other social network characteristics in future models. The most important future direction is to use machine learning models that track and measure opinions changing over time, e.g. using recurrent models like Long Short Term Memory applied at the user level. These improvements are necessary for better election predictor models.

We may expect that in the future traditional surveys may be incrementally replaced by these new non-intrusive methods. AI is a thermometer that provides the key to predicting not only elections but the great trends that develop at the local and global levels in society. We have shown how AI allows to synthesize the opinion of millions of people including those silent majorities of hidden voters who would not be heard otherwise. We must not ignore that people are tired of answering surveys. AI can then deduce, predict, interpret and understand what people want to express.

Abbreviations

PASO: In Spanish: Primarias, Abiertas, Simultáneas y Obligatorias; in English: Open, Simultaneous and Obligatory Primaries; AF: Alberto Fernández; MM: Mauricio Macri; CFK: Cristina Fernández de Kirchner; FF: Fernández-Fernández; MP: Macri-Pichetto; TP: Third party; LR: Logistic regression; SVM: Support vector machine; NB: Naive bayes; RF: Random forest; DT: Decision tree.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40537-021-00525-8.

Additional file 1: 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. **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 classifiedusers for each candidate. **Fig. S1.** Instantaneous prediction compared with trusted polls. Thick lines represent AI prediction and dashed line represent trusted polls. We use the 7, 14,21, 28-day window 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 instantaneous opinion model. **Fig. S2.** Cumulative prediction with T0=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 (T0=Mar, Apr,May, Jun, Jul, Aug 1). Therefore, we select March 1 as the initial date with the least error. **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. **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.

Authors' contributions

HAM conceived of and designed the research. ZZ and MS conducted the analysis and analyzed the results. LC conducted the polls for Elypsis before and after the Argentina 2019 election, and analyzed the Elypsis polls results. ZZ, MS, GC, and HAM wrote the paper. All authors read and approved the final manuscript.

Funding

Partial support was provided by NIH R01 EB028157. Z.Z. acknowledges support from Special Fund for Fundamental Scientific Research of the Beijing Colleges in CUEB. G.C. acknowledges support from EU project, HUMANE-AI-NET (grant number 952026). H.A.M. owns shares of Kcore Analytics.

Availability of data and materials

We have made the raw data and the codes publicly available at https://osf.io/z7j4r/ and https://github.com/MakseLab/TwitterElectionPredictor (see also shorturl.at/hkBF3). The results presented in this paper can be reproduced with the code and data provided in these links.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]School of Statistics, Capital University of Economics and Business, 100070 Beijing, China. [2]IMT School for Advanced Studies, 55100 Lucca, Italy. [3]Seido, Moldes 2277, Buenos Aires, Argentina. [4]Department of Molecular Sciences and Nanosystems, Ca' Foscari University of Venice, 30172 Venice, Italy. [5]European Centre for Living Technology, Italy, Venice, Italy. [6]Institute for Complex Systems, Consiglio Nazionale delle Ricerche, UoS Sapienza, 00185 Rome, Italy. [7]London Institute for Mathematical Sciences, W1K2XF London, United Kingdom. [8]Levich Institute and Physics Department, City College of New York, 10031 New York, USA.

Received: 8 July 2021 Accepted: 4 October 2021
Published online: 23 October 2021

References

1. Tourangeau R, Conrad FG, Couper MP. The science of web surveys. New York: Oxford University Press; 2013.
2. Kennedy C, Blumenthal M, Clement S, Clinton JD, Durand C, Franklin C, McGeeney K, Miringoff L, Olson K, Rivers D, et al. An evaluation of the 2016 election polls in the United States. Public Opin Q. 2018;82(1):1–33.
3. Durand C, Blais A. Quebec 2018: a failure of the polls? Can J Polit Sci/Revue Canadienne de Science Politique. 2020;53(1):133–50.
4. Duncan P, The Guardian. How the pollsters got it wrong on the EU referendum. 2016. https://www.theguardian.com/politics/2016/jun/24/how-eu-referendum-pollsters-wrong-opinion-predict-close. Accessed 14 Oct 2021.
5. Cohn N. The Upshot, New York Times. Why Polls Have Been Wrong Recently. 2016. https://www.nytimes.com/2016/01/08/upshot/why-polls-have-been-wrong-recently.html. Accessed 14 Oct 2021.
6. Jacobs J, House B. Trump says he expected to lose election because of poll results. Boomberg Politics;2016.
7. Kennedy C, Hartig H. Response rates in telephone surveys have resumed their decline. Pew Research Center ;2019.
8. Battaglia MP, Izrael D, Hoaglin DC, Frankel MR. Tips and tricks for raking survey data (aka sample balancing). Abt Assoc. 2004;1:4740–4.
9. Izrael D, Hoaglin DC, Battaglia MP. A sas macro for balancing a weighted sample. In: Proceedings of the Twenty-fifth Annual SAS Users Group International Conference, pp. 9–12 ;2000. Citeseer.
10. Leonhardt D. New York Times. 'A Black Eye': why political polling missed the mark. Again. 2020. https://www.nytimes.com/2020/11/12/us/politics/election-polls-trump-biden.html. Accessed 14 Oct 2021.
11. Payne JG. The Bradley effect: mediated reality of race and politics in the 2008 US presidential election. Am Behav Sci. 2010;54(4):417–35.
12. Krumpal I. Determinants of social desirability bias in sensitive surveys: a literature review. Qual Quant. 2013;47(4):2025–47.
13. Zolghadr M, Niaki SAA, Niaki S. Modeling and forecasting us presidential election using learning algorithms. J Ind Eng Int. 2018;14(3):491–500.
14. Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Based Systems. 2015;89:14–46.

15. Jaidka K, Ahmed S, Skoric M, Hilbert M. Predicting elections from social media: a three-country, three-method comparative study. Asian J Commun. 2019;29(3):252–73.
16. Jungherr A. Twitter use in election campaigns: a systematic literature review. J Inf Technol Polit. 2016;13(1):72–91.
17. Bovet A, Morone F, Makse HA. Validation of twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. Sci Rep. 2018;8(1):1–16.
18. Bovet A, Makse HA. Influence of fake news in twitter during the 2016 US presidential election. Nat Commun. 2019;10(1):1–14.
19. Papakyriakopoulos O, Hegelich S, Shahrezaye M, Serrano JCM. Social media and microtargeting: political data processing and the consequences for Germany. Big Data Soc. 2018;5(2).
20. Tumasjan A, Sprenger T, Sandner P, Welpe I. Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Proceedings of the International AAAI Conference on Web and Social Media, 2010;vol. 4.
21. Jungherr A, Jürgens P, Schoen H. Why the pirate party won the German election of 2009 or the trouble with predictions: a response to Tumasjan, A., Sprenger, TO, Sander, PG, & Welpe, IM "Predicting elections with twitter: what 140 characters reveal about political sentiment". Soc Sci Comput Rev. 2012;30(2):229–34.
22. Gaurav M, Srivastava A, Kumar A, Miller S. Leveraging candidate popularity on twitter to predict election outcome. In: Proceedings of the 7th workshop on social network mining and analysis, 2013;1–8.
23. Lui C, Metaxas PT, Mustafaraj E. On the predictability of the US elections through search volume activity. http://repository.wellesley.edu/scholarship/23/. Accessed 14 Oct 2021.
24. Bermingham A, Smeaton A. On using twitter to monitor political sentiment and predict election results. In: Proceedings of the workshop on sentiment analysis where AI meets psychology (SAAIP 2011), 2011;2–10.
25. Ceron A, Curini L, Iacus SM, Porro G. Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France. New Media Soc. 2014;16(2):340–58.
26. Caldarelli G, Chessa A, Pammolli F, Pompa G, Puliga M, Riccaboni M, Riotta G. A multi-level geographical study of Italian political elections from twitter data. PloS One. 2014;9(5):95809.
27. Singh P, Sawhney RS, Kahlon KS. Forecasting the 2016 us presidential elections using sentiment analysis. In: Conference on e-Business, e-Services and e-Society, 2017; 412–423 . Springer.
28. Xia E, Yue H, Liu H. Tweet sentiment analysis of the 2020 US presidential election. In: Companion Proceedings of the Web Conference 2021, 2021;367–371.
29. Singh P, Dwivedi YK, Kahlon KS, Pathania A, Sawhney RS. Can twitter analytics predict election outcome? an insight from 2017 Punjab assembly elections. Gov Inf Q. 2020;37(2):101444.
30. Budiharto W, Meiliana M. Prediction and analysis of Indonesia presidential election from twitter using sentiment analysis. J Big Data. 2018;5(1):1–10.
31. Newman M. Networks: an introduction. New York: Oxford University Press; 2010.
32. Cuzzocrea A, Papadimitriou A, Katsaros D, Manolopoulos Y. Edge betweenness centrality: a novel algorithm for qos-based topology control over wireless sensor networks. J Netw Comput Appl. 2012;35(4):1210–7.
33. Bode L, Dalrymple KE. Politics in 140 characters or less: campaign communication, network interaction, and political participation on twitter. J Polit Market. 2016;15(4):311–32.
34. Paladini E. Encuestadoras bajo fuego: por qué erraron en las PASO y qué dicen para octubre 2019. https://www.clarin.com/politica/encuestadoras-fuego-erraron-paso-dicen-octubre_0_T72H9hdl.html. Accessed 14 Oct 2021.
35. Jasny BR, Stone R. Prediction and its limits. Science. 2017;355:468–9.
36. Wikipedia. https://es.wikipedia.org/wiki/Anexo:Encuestas_de_intencion_de_voto_para_las_elecciones_presidenciales_de_Argentina_de_2019. Accessed 14 Oct 2021.
37. Bonelli M. https://www.clarin.com/opinion/intrigas-casa-rosada-pases-factura-city-lunes-negro_0_jnggAlsh5.html. Accessed 14 Oct 2021.
38. Levy R. Wall Street Journal. 2019. https://www.wsj.com/articles/hedge-fund-loses-1-billion-in-one-month-on-argentina-bet-11567696547. Accessed 14 Oct 2021.
39. Manning C, Schutze H. Foundations of statistical natural language processing. MA, New York: Cambridge; 1999.
40. Deng L, Liu Y. Deep learning in natural language processing. Cham, Switzerland: Springer; 2018.
41. Martinez-Romo J, Araujo L, Borge-Holthoefer J, Arenas A, Capitán JA, Cuesta JA. Disentangling categorical relationships through a graph of co-occurrences. Phys Rev E. 2011;84(4):046108.
42. Ceron A, Curini L, Iacus SM. Using sentiment analysis to monitor electoral campaigns: method matters-evidence from the United States and Italy. Soc Sci Comput Rev. 2015;33(1):3–20.
43. An J, Weber I. #greysanatomy vs #yankees: Demographics and hashtag use on Twitter. In: Proceedings of the International AAAI Conference on Web and Social Media; 2016. p. 10.
44. Vikatos P, Messias J, Miranda M, Benevenuto F. Linguistic diversities of demographic groups in Twitter. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media; 2017. p. 275–84.
45. New York Times National Polling Average. The Upshot. http://www.nytimes.com/interactive/2016/us/elections/polls.html. Accessed 14 Oct 2021.
46. SEIDO - Special Report: Lie to Me. https://us3.campaign-archive.com/?e=&u=e02ede36ce39515be5fb17728&id=3bf5cf2e90. Accessed 14 Oct 2021.

Publisher's Note