YouTube Recommendations Do Predict Polls: A note on the 2022 French presidential election

Erwan Le Merrer, Gilles Trédan, Ali Yesilkanat

YouTube Recommendations Do Predict Polls:
A note on the 2022 French presidential election

Erwan Le Merrer (Inria), Gilles Trédan (LAAS/CNRS) and Ali Yesilkanat (Inria)

Abstract

We ask if YouTube political recommendations follow the polls performed during the 2022 presidential campaign in France. We find that these recommendations to users, particularly the exposure time in recommended videos, allow for a good prediction of polls, with a daily Mean Absolute Error of solely 3.24% over all candidates. Compared to final election results, polls were 1.11% accurate, whereas our approach was 1.93% accurate.

Close to 50M people were accessing YouTube every month in France in 2021 (statistics by BDM). We wonder if this mass access constitutes a sufficient signal to estimate another signal captured from the public at the same period: polls.

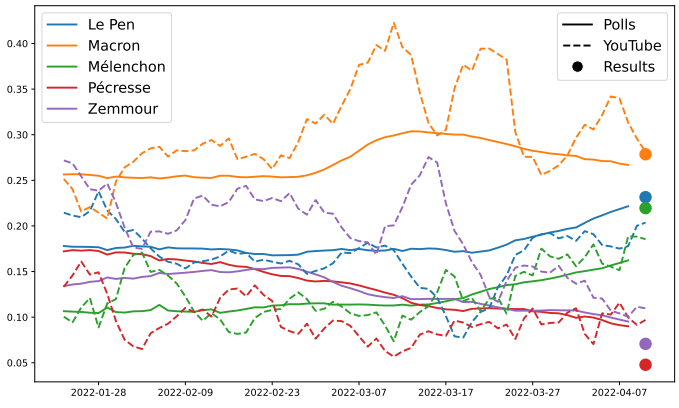
Methodology We consider the twelve candidates able to run for the presidency officially. We set up automated scripts that simulate users watching videos on YouTube. At every simulation, a cookie-free fresh bot goes on the French "National News" category, watches a random video, and the 4 following autoplay videos.. This action is performed around 180 times per day, from January 17th to April 10th (first-round election day). In this experiment, we extract the transcripts of the 5 video sequences and search for candidate names within each transcript. The time duration of a sentence in which a candidate is mentioned is counted as exposure time to her credit. We aggregate the total exposure time of each candidate over the course of a day and normalize this value by the total exposure time of all candidates: this yields a ratio representing each candidate's exposure timeshare (ETS). This value is directly compared the polls collected and made available by the Pollotron website (please refer to the Appendix).

Results Figure (a) presents both polls (solid curve, y-axis) and measured ETS (dashed curve, y-axis) for the top-5 highest candidates over 3 months preceding the first election round (x-axis); both are smoothed (7-day rolling window).

While ETS values are less stable than polls, both generally exhibit a close match throughout the period. For instance, this statement should be nuanced for some candidates, with Zemmour being systematically over-rated by ETS and Le Pen conversely under-rated. Interestingly, both polls and ETS provide a good estimation of candidates' actual results during the first election round (represented as dots), respectively exhibiting average errors of 1.11% and 1.93%.

Table (b) presents the Mean Absolute Error between ETS and polls for the 12 candidates over the whole monitored period. We also report the average exposure time per candidate that our scripts encountered for relative comparison.

We find this 3.24% prediction to be very significant, with the notable remark that ETS ends up with less than 1% error more than the last polls before the results. As polls rely on considerably fewer people than the scale at which people impact recommendations, this is certainly an interesting research track for future events.



(a) Evolution of polls and of the ETS metric over the campaign, for the top-5 highest-scored candidates.

Candidates	MAE (%)	Avg ETS / day (s)	Election Results (%)
Arthaud	0.30	9.48	0.56
Dupont-Aignan	1.51	3.41	2.06
Hidalgo	1.01	63.92	1.74
Jadot	3.14	71.57	4.63
Lassalle	1.00	23.74	3.13
Le Pen	8.70	544.08	23.15
Macron	3.26	773.99	27.85
Mélenchon	3.79	378.47	21.95
Poutou	0.53	15.81	0.76
Pécresse	8.02	213.76	4.78
Roussel	2.26	65.42	2.28
Zemmour	5.34	424.57	7.07
Avg	**3.24**	**215.68**	N/A

(b) Mean Absolute Error (MAE) between polls and our Exposure Time Share (ETS) measure on first column. Average exposure time witnessed by our scripts in a day.

Appendix: recommendations as a predictor

Extracting exposure time from videos Let V a set of videos. Given $v \in V$, by parsing its transcript, we extract (an approximation of) each candidate's exposure time by counting the duration (in seconds) of sentences in which this candidate's name appears. Let $t_{v,c}$ this value.

Runs collecting videos We did perform an average of 180 script runs per day. Let t be the day, let W_t the set of runs of that day. A run $w \in W_t$ consists of a random click on a video v_1 of the National news page, followed by the autoplayed (first suggested video played by default after current video is over) $v_2 = a(v_1)$, followed by the autoplayed $v_3 = a(v_2)$, and so on and so forth, until v_5. We conveniently write $v(w) = \{v_1, \ldots, v_5\}$ to designate the set of videos returned by a run w. Similarly, we define $v(W_t) = \cup_{w \in W_t} v(w)$ to be the videos returned by runs over day t.

Candidate Daily Exposure Time Let $s_c(t)$ be the total exposure time of candidate c found in videos fetched on day t, which we define as $et_c(t) = \sum_{v \in v(W_t)} t_{v,c}$. From there, we define the *Exposure Time Share (ETS)* of candidate c on day t as: $r_c(t) = \frac{et_c(t)}{\sum_c et_c(t)}$.

Polls and matching metric Over the measurement period, we collect the aggregated poll values from Pollotron[1]. Let $pol_c(t)$ be the poll score of candidate c on day t.

To measure the agreement of polls and ETS, we rely on standard Mean Average Error. Given a candidate c, we define it as the absolute difference between predicted (ETS) value and poll value, averaged over all days of the measurement period T. Formally: $\text{MAE}_c = \frac{1}{T} \sum_t |r_c(t) - pol_c(t)|$.

This MAE approach is also used to compare ETS and polls against the actual results obtained by each candidate at the election first round. First, let res_c be the result (i.e. the fraction of valid expressed votes) obtained by candidate c. Let t_f be the election day, and t'_f be the last polling day[2]. The average error between ETS (resp. polls) and results is computed as: $err_{ETS} = \frac{1}{12} \sum_c |r_c(t_f) - res_c|$, resp. $err_{POL} = \frac{1}{12} \sum_c |pol_c(t'_f) - res_c|$. A value of 0.012 can be interpreted as: "on average, predictions were 1.2% off the actual result".

[1] `https://datapolitics.fr/agregateur-sondages-presidentielle2022/`

[2] French legislation imposes polls to stop 24 hours before the vote day.