Les Recommandations de YouTube Prédisent les Sondages:
Une note sur l'élection présidentielle Française de 2022

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

Abstract

Nous nous demandons si les recommandations politiques de YouTube suivent les sondages réalisés lors de la campagne présidentielle de 2022 en France. Nous constatons que ces recommandations aux utilisateurs, notamment le temps d'exposition des candidats dans les vidéos recommandées, permettent une bonne prédiction des sondages, avec une erreur moyenne absolue quotidienne (MAE) de seulement $3,24\%$ sur l'ensemble des candidats. Par rapport aux résultats finaux des élections, les sondages ont présenté une erreur de $1,11\%$, alors que celle de notre approche a été de $1,93\%$.

Près de 50M de personnes accédaient à YouTube chaque mois en France en 2021 (stat. BDM). Nous postulons que cet accès massif constitue un signal en relation avec signal capté auprès du public à la même période : les sondages.

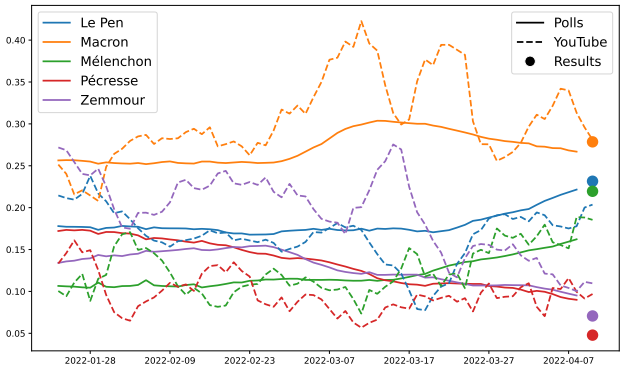
Méthodologie Nous prenons en compte les douze candidats qui furent en mesure de se présenter officiellement pour la campagne. Nous avons mis en place des scripts automatisés (bots) qui simulent des utilisateurs regardant des vidéos sur YouTube. A chaque simulation, un nouveau bot sans cookie se rend sur la catégorie française "Actualités nationales", regarde une vidéo aléatoire, et les 4 vidéos suivantes proposées en *autoplay*. Cette action est effectuée environ 180 fois par jour, du 17 janvier au 10 avril (jour du premier tour des élections). Nous extrayons les transcriptions des 5 vidéos ainsi vues, et recherchons les noms des candidats dans chacune. La durée d'une phrase dans laquelle un candidat est mentionné est comptée comme temps d'exposition et mise à son crédit. Nous agrégeons le temps d'exposition total de chaque candidat au cours d'une journée et normalisons cette valeur par le temps d'exposition total de tous les candidats : nous obtenons ainsi un ratio représentant le temps d'exposition partagé (ETS) de chaque candidat. Cette valeur est directement comparée aux sondages mis à disposition par le site Pollotron (voir en annexe).

Résultats La figure (a) présente à la fois les sondages (axe y) et les valeurs d'ETS (courbes en pointillés, axe y) pour les cinq premiers candidats les plus en vue au cours des trois mois précédant le premier tour des élections (axe x); les courbes sont lissées (fenêtre glissante de 7 jours).

Bien que les valeurs ETS soient moins stables que les sondages, les deux présentent généralement une correspondance étroite tout au long de la période. Cette affirmation doit être nuancée pour certains candidats, Zemmour étant systématiquement surévalué par l'ETS et Le Pen inversement sous-évaluée. Il est intéressant de noter que les sondages et l'ETS fournissent tous deux une bonne estimation des résultats réels des candidats lors du premier tour de l'élection (représentés par des points), présentant respectivement des erreurs moyennes de $1,11\%$ et $1,93\%$.

Le tableau (b) présente l'erreur absolue moyenne (MAE) entre les ETS et les sondages pour les 12 candidats sur l'ensemble de la période. Nous indiquons également le temps d'exposition moyen par candidat que nos scripts ont rencontré pour une comparaison relative.

L'erreur moyenne de prédiction de $3,24\%$ est remarquablement basse, et les ETS finaux prédisent les résultats des élections avec moins de 1% d'erreur de plus que les derniers sondages. Étant donné que les sondages reposent sur un nombre de personnes considérablement inférieur à l'échelle à laquelle les personnes ont un impact sur les recommandations, il s'agit certainement d'une piste de recherche intéressante pour les événements futurs.



(a) Évolution des sondages (Polls) et de l'ETS (YouTube) sur la campagne, pour les 5 candidats les mieux placés.

Candidats	MAE (%)	ETS moy. / jour (s)	Résultats (%)
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
Moyenne	**3.24**	**215.68**	N/A

(b) Erreur moyenne absolue (MAE) entre les sondages et l'ETS sur toute la période, en première colonne. Temps moyen d'exposition mesuré par nos scripts par jour, et résultats finau du 10 avril.

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's first round. First, let res_c be the result (i.e. the fraction of validly 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 are 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 voting day.