#Élysée2017fr: The 2017 French Presidential Campaign on Twitter

Ophélie Fraisier,[1,2] Guillaume Cabanac,[2] Yoann Pitarch,[2]
Romaric Besançon,[3] Mohand Boughanem[2]
[1]CEA-Tech Occitanie, Toulouse, France
[2]IRIT, Université de Toulouse, CNRS, Toulouse, France
[3]CEA LIST, Nano-INNOV, Palaiseau, France
{first name.last name}@{cea, irit}.fr

Abstract

The French presidential election was one of the main political event of 2017, and triggered a lot of activity on Twitter. The campaign was highly unpredictable and led to the rise of 5 main parties instead of the historical bipartite (left-right) confrontation, ranging from far-left to far-right. This dataset paper proposes #Élysée2017fr, a large and complex dataset of 22,853 Twitter profiles active during the campaign (from November 2016 to May 2017), and their corresponding tweets and retweets, plus the retweet and mention networks related to these profiles. The profiles were manually annotated with their political affiliations (up to 2 political parties per profile), their nature (individual or collective), and the sex of the profile's owner when available. This is one of the rare datasets that considers a non-binary stance classification and, to our knowledge, the first one with a large number of profiles, and the first one proposing overlapping political communities. This dataset can be used as-is to study the campaign mechanisms on Twitter, or used to test stance detection models or network analysis tools. Mining these data might reveal new insights on current issues like echo chambers or fake news diffusion.

1 Introduction

Twitter became a key research support for computer scientists and social scientists alike these past few years, due to its wide use and the ease in collecting its data. To quote Golder and Macy (2015), "Twitter has emerged as the single most powerful "socioscope" available [...] for collecting fine-grained time-stamped records of human behavior and social interaction". It has been particularly used for studying political topics such as elections or public reactions on particular policies. Several works focus on predicting the political stance of Twitter profiles, by using retweets, hashtags, or following behavior (Makazhanov and Rafiei 2013; Conover et al. 2011b; Weber, Garimella, and Teka 2013; Pennacchiotti and Popescu 2011). These works are often closely linked with research on polarization and the "echo chambers" popularized by Sunstein (Conover et al. 2011a; Weber, Garimella, and Batayneh 2013; Fraisier et al. 2017). Others strive to measure the predictive power of tweets for elections results, a topic counting for the time being as much enthusiasts (Tumasjan et al. 2010; Sang and Bos 2012) as detractors (Gayo-Avello 2012; Jungherr, Jürgens, and Schoen 2012). Another area of research is of course social media abuse, with for example smear campaigns and "fake news" (Ratkiewicz et al. 2011; Saez-Trumper 2014; Vosoughi, Mohsenvand, and Roy 2017).

While these works tackle various compelling research questions, they all require an annotated dataset of Twitter profiles as input. Unfortunately, high-quality annotated datasets are a rare commodity, despite being essential for improving and reliably measuring models performances. While collecting data is easy enough nowadays, annotating a dataset is a harsh task, which explains why existing datasets often offer a small quantity of annotations (Kratzke 2017), or focus on binary situations, opposing for example Democrats and Republicans, or "No" and "Yes" partisans in the 2014 Scottish independence referendum (Brigadir, Greene, and Cunningham 2015).

We propose in this dataset paper #Élysée2017fr: an original, large and complex dataset focused on the 2017 French presidential election. The main characteristics of this dataset are:

- 22,853 Twitter profiles engaging on a major French political event, manually annotated by experts.
- 6 political stances (5 parties and an *undefined* category).
- Overlapping affiliations to political parties.
- Supplementary data concerning the nature of the profiles and the sex of the owners when available.
- Ids for 2,414,584 tweets and 7,763,931 retweets discussing this election in several languages.
- Retweet and mention networks.

This dataset was built to propose a study case that is closer to real-life complex situations than existing datasets. It can be used as-is by political scientists for studying the campaign mechanisms on Twitter, and thanks to the manual annotations, it can also be used by computer scientists to test stance detection models or network analysis tools. In addition, its large size guarantees a better robustness for supervised models, and the presence of overlapping political communities helps with the exploration of advanced research questions, such as the identification of swing-voters. Section 2 presents the existing datasets focused on user-level

Parties	Extracts of profile's Twitter biographies	
PS / EM	Fidelity @fhollande support #EnMarche [fr] Fidélité @fhollande soutien #EnMarche	"@fhollande" references François Hollande, previous French socialist president.
LR / FN	#FillonistWithMarine [. . .] #NeverMacron [fr] #FillonisteAvecMarine [. . .] #JamaisMacron	"Marine" references Marine Le Pen, leader of FN.
EM / LR	#EnMarche Ex-Young with Juppé [fr] #EnMarche Ex-Jeune avec Juppé	"Juppé" references Alain Juppé, one of the leaders of LR who was eliminated during the right-wing primaries.

Table 1: Examples of Twitter profiles affiliated to several parties during the 2017 French presidential campaign.

political stance, while Section 3 delves into the political context surrounding this new dataset. Its harvesting and annotation process are detailed in Section 4. Section 5 describes the annotated data, and Section 6 raises some of its limitations. Finally, Section 7 presents in more detail several possible use cases.

2 Related datasets

Several datasets concerning user-level political stance appeared in the recent litterature. These datasets cover various topics, but while they are precious tools to analyse political discourse online, a number of them uses automatic methods to determine profiles' stances, usually Bayesian or graph-based models built on retweets or follow graphs. The absence of manual verification means that these datasets cannot be reliably used to improve and test new stance detection models since the errors cannot be reliably accounted for. Among the topics covered by automatically annotated datasets are the 2009 German federal election (Jungherr 2013; Tumasjan et al. 2010), the 2011 Dutch senate election (Sang and Bos 2012), the 2011 Spanish legislative elections, the 2012 and 2016 US presidential elections, Brexit, ObamaCare, abortion policies, or fracking policies (Barberá and Rivero 2015; Garimella et al. 2017).

Despite being scarce, some manually annotated datasets do exist:

- Brigadir, Greene, and Cunningham (2015) collected datasets on the 2014 Scottish Independence Referendum and the 2014 US midterms elections. They each have two opposing viewpoints, and contain respectively 1,218 and 1,939 profiles which were manually selected by experts. The main drawback of these datasets is the nature of the selected profiles: they are mainly politicians or high-profile activists, and therefore not representative of "standard" Twitter profiles discussing political topics.
- Kratzke (2017) collected the tweets published by 364 profiles of German politicians during the 2017 German federal election. The profiles were manually selected and can belong to one of the 6 considered parties. This dataset offers non binary stances, but is very small and, like the aforementioned datasets, focused of politicians profiles.
- Lu, Caverlee, and Niu (2015) shared manually annotated datasets of 504 profiles on gun control, abortion, and ObamaCare. The profiles are categorized on a 5-point scale going from "strong support" to "strong opposition". While the engagement scale is interesting, the datasets are small, and the problematics remain presented as an opposition of two main stances.
- Preoţiuc-Pietro et al. (2017) built a dataset concerning US politics of 3,938 profiles categorized on a 7-point scale going from "very conservative" to "very liberal" labels self reported through surveys. The question remains framed as a binary confrontation between "conservative" and "liberal", but the self reporting of profiles' political leaning is an undeniable guarantee of quality, mitigating the medium size of the dataset.

As deducible from the above presentations, the existing datasets have difficulty capturing the complexity and disorder of real-life situations. They usually focus on a small number of binary annotations, or a specific category of profiles, and cannot take into account a multiplicity of viewpoints having varying proximities with each other. We provide a new large dataset, manually annotated by experts, and featuring potentially *overlapping* stances in order to capture undecided profiles or profiles ideologically close to several parties. It aims to enable computer scientists to develop and refine their models using high quality manual annotations, and political scientists to study the French presidential elections presence on Twitter, in line with existing studies on European elections (Jungherr 2013; Tumasjan et al. 2010; Sang and Bos 2012; Brigadir, Greene, and Cunningham 2015; Kratzke 2017).

3 Political context the dataset captures

The dataset was collected during the 2017 French presidential campaign which ended May 7th with the election of current French president, Emmanuel Macron. Instead of the usual scenario of a presidential elections dominated by the candidates of the historical left- and right-leaning parties, this campaign was highly atypical and unpredictable, with moving allegiances shifting around the five main parties, which makes it a perfect ground for stance analysis and stance detection.

Before the campaign, primaries were organised by the parties in order to select a right candidate and a left candidate. While these primaries resulted in candidate nominations, namely *François Fillon* for the right-leaning party *Les Républicains* and *Benoît Hamon* for the left-leaning *Parti Socialiste*, they also led to a lot of strife on both sides. The situation was made even more unstable by the presence of a popular newly created party, *En Marche*, *Emmanuel Macron*'s movement, introduced as a movement uniting both

Party	Individual (# media professionals)						Non individual		Total	# Tweets	# Retweets
	Male		Female		Other / Und.		Political	Other			
FI	2696	(12)	1093	(1)	1023	(2)	298	3	**5,113**	528,401	1,664,144
PS	697	(1)	610	(2)	322		200	3	**1,832**	162,368	533,468
EM	2056	(2)	766	(8)	473	(3)	654	13	**3,962**	379,223	1,142,000
LR	2224	(3)	922	(2)	606	(1)	604	10	**4,366**	544,259	2,136,897
FN	1878	(4)	567	(2)	812	(1)	114	5	**3,376**	330,614	1,525,264
FI / PS	91	(1)	67		68		2		**228**	20,389	29,957
FI / EM	19		9		5				**33**	2,987	6,180
FI / LR	3				1				**4**	426	323
FI / FN	14		2		6				**22**	3,804	4,836
PS / EM	84		43		22		2		**151**	21,557	70,449
PS / LR					2				**2**	353	1,309
PS / FN	2				1				**3**	1,760	9,410
EM / LR	86		34		18		9	1	**148**	24,504	37,303
EM / FN	1								**1**	1,070	1,459
LR / FN	113	(1)	37		56		4	1	**211**	24,186	113,995
Und.	1428	(198)	700	(92)	1049	(28)	35	189	**3,401**	368,683	486,937
Total	**11,392**	(222)	**4,850**	(107)	**4,464**	(35)	**1,922**	**225**	**22,853**	**2,414,584**	**7,763,931**
Off topic									316		

Table 2: Number of profiles by political party.

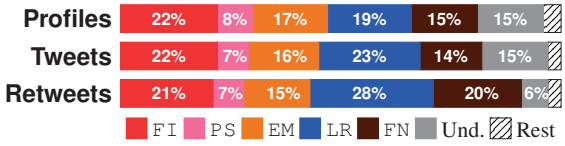

Figure 1: Share of profiles, tweets, and retweets by party.

the left and the right. The extremes were also very active in this campaign. While the *Front National*, France's main far-right movement led by *Marine Le Pen*, had been a strong contender for the past 2 presidential elections, this last campaign saw the emergence of a large far-left movement called *La France Insoumise*, led by *Jean-Luc Mélenchon*. The campaign was further shaken by revelations of allegedly fictitious jobs held by Penelope Fillon, François Fillon's wife, and their children. While François Fillon was given favourite when this affair arose, his decision to maintain his candidature despite his placement under formal investigation led to further dissension, with some historical figures of the party and electors choosing to report their vote on Marine Le Pen or Emmanuel Macron instead.

4 Study design

4.1 Profile selection

The dataset was built from November 25th 2016 to May 12th 2017, by monitoring several keywords referencing the political parties and candidates involved using DMI-TCAT (Borra and Rieder 2014). These keywords were selected by researchers familiar with the French political landscape on Twitter, and are detailed in the file `keywords.csv` shared with the dataset. This monitoring resulted in a dataset of 42,251,431 tweets published by 2,941,991 profiles. Given the size of the dataset, annotating all the profiles was intractable and several choices were made:

1. Determining the political stance of someone based on his / her publications is not an easy task, even for a human, and particularly on Twitter where the difficulty increases due to the publications' brevity. In order to obtain reliable annotations, we only kept those profiles featuring one of the five main parties in the presentation area. By "*profile presentation*" we mean all the *pseudonyms* (also known as Twitter handles), *names*, and *biographies* used by a profile during the campaign.
2. The profiles having less than 10 posts (including original tweets, and retweets) during the monitored period were also discarded.

The resulting subset of 23,169 profiles was then manually annotated. While this represents only 1% of the profiles present in the original dataset, it still is a significantly larger amount of annotations than the existing datasets.

4.2 Annotations

As presented in Section 3, 2017 French presidential campaign was led by 5 main parties, presented here from far-left to far-right:

FI La France Insoumise [Unbowed France], led by Jean-Luc Mélenchon,

PS Le Parti Socialiste [Socialist Party], led by Benoît Hamon,

EM En Marche ! [Forward! or Working!], led by Emmanuel Macron,

LR Les Républicains [Republicans], led by François Fillon,

FN Le Front National [National Front], led by Marine Le Pen.

Given the unstable nature of this campaign, some profiles can be attached to several parties instead of a unique classification. Some examples are given in Table 1.

For the annotation process, we involved domain experts instead of crowdsourced workers. Indeed, in order to obtain quality annotations, we needed people understanding French, and having a good knowledge of the French political landscape and the campaigns' events. They had to determine political affiliations based on the support explicitly expressed for parties or candidates. An official affiliation (being a party member for example) was not required to consider that a profile was affiliated to a party, and a "default" vote was not considered a genuine affiliation (for example a tweet indicating "I vote against Le Pen with a Macron bulletin" is not considered a source of affiliation to EM).

In order to facilitate the task for the 16 annotators, each profile was shown as its presentation plus its 10 most shared publications (including original tweets, and retweets). If these elements were not precise enough to categorize a profile, it was classified as having undetermined political preferences. The annotation process was divided into 2 steps:

- We first selected 1,000 profiles at random to be annotated by three different annotators to measure the inter-annotator agreement. While many inter-annotator agreement measures have been used over the years (Artstein and Poesio 2008), our annotation scheme required a non-trivial one. Fleiss' Kappa (1971) being designed for a single-categorisation annotation, we opted for Bhowmick, Mitra, and Basu's variant (2008), which is specifically designed to compute agreement when items can be categorized into more than one class.[1] For these 1,000 profiles, the annotations resulted in an observed agreement $P_o = 0.89$, a chance agreement $P_e = 0.57$, and an agreement measure $A_m = 0.75$, indicating a strong consensus between annotators. The final annotations were obtained by majority voting.

- Given the inter-annotator agreement measured in the previous step, and the size of our dataset, we chose to annotate the rest of the corpus with only one annotator per profile.

In addition to the political affiliations of the profiles, we annotated the following information:

- The nature of the profile's ownership: an individual, or a group or entity.

- For profiles owned by an individual: the sex of the profile's owner when it is easily deducible, and if the owner self-identify as a media professional.

- For other profiles, if the group or entity running the profile is of political nature (for example official profiles of the parties, or profiles of activist groups).

5 Description

This section presents some descriptive and quantitative analysis of this dataset. The formats and technical details of the files published along this paper are given in Appendix A.

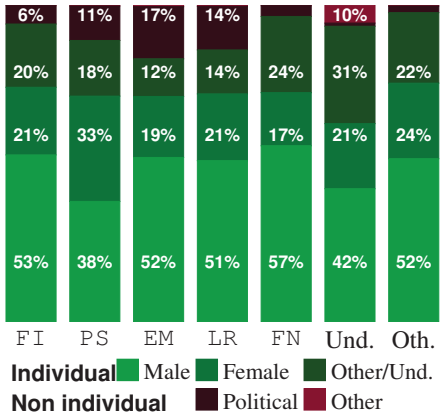

Figure 2: Distribution of profiles nature by party.

Table 3: Median (med) and average (avg) number of publications by profile and by party.

		FI	PS	EM	LR	FN	Und.
Tweets	Med	23	25	25	28	29	15
	Avg	107	93	99	133	106	118
Retweets	Med	46	77	54	66	98	26
	Avg	330	293	292	496	457	142

5.1 Profiles and publications by party

Table 2 presents the number of profiles by party, as well as the volume of tweets and retweets they published, while Figure 1 compares the proportion of profiles, tweets, and retweets each party represents. In terms of profiles, the most represented parties are FI and LR, while PS is largely distanced by the other parties. The proportions are similar for tweets, while for retweets LR represents almost a third of all publications and FN is similar to FI, suggesting that LR and FN profiles are particularly active. Indeed, they respectively account for 28% and 20% of retweets while representing only 19% and 15% of the total number of profiles. This is confirmed by Table 3, with a median of 66 retweets by profile for LR and 98 retweets for FN. Interestingly, the distribution for PS retweets also suggests highly active users, with a median of 77 retweets by profile, but this is not visible in the total volume of publications. This table also confirms that highly active users in terms of tweet publications are rare, while highly active retweeters are much more common, since the median number of tweet publication ranges from 15 to 29, against a median number of retweets ranging from 26 to 98.

Figure 2 compares the distribution of profile nature by party. In terms of individual profiles, it shows no significant difference between parties, with a large majority of men running the profiles. However, EM, LR, and PS have noticeably

[1]The Python implementation of this measure is available at the following address: https://github.com/SyrupType/bhowmick_agreement_measure

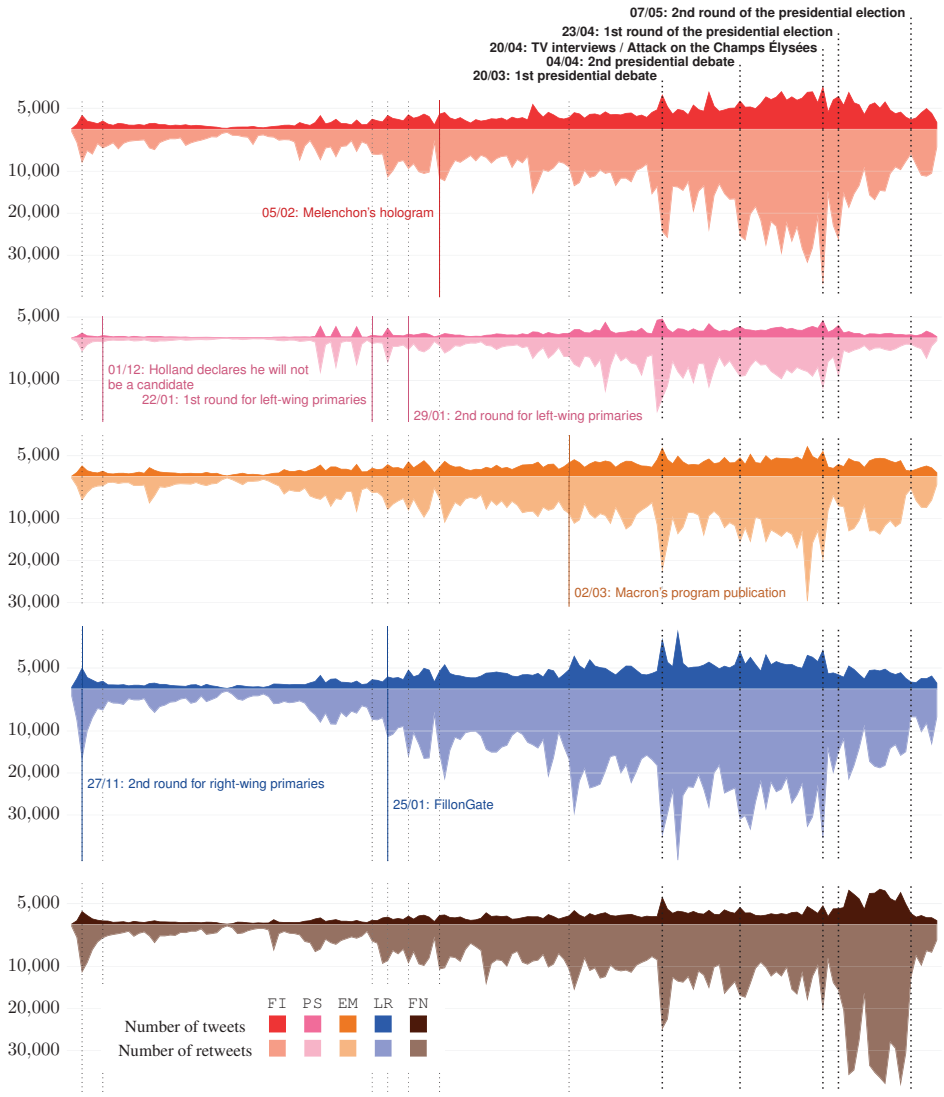

Figure 3: Evolution of the number of tweets and retweets published by annotated profiles between November 25th 2016 and May 12th 2017.

more profiles owned by activist groups or associations. This is interesting since, while for LR and PS it can result from the seniority of the parties, allowing for activist groups to be already prepared and present on Twitter, EM was formed a few months prior to the campaign. This could represent an active effort by the party to be organized and present on the platform.

5.2 Temporal analysis

Timeline. Figure 3 shows the evolution of the number of tweets and retweets published by the profiles in our dataset. For the sake of simplicity, this analysis is focused on the five main parties, with the publications of profiles having several political affiliations being evenly distributed among them. The main peaks of activity correlate with campaign events, particularly the debates and rounds of the election (some important campaign events are reported in Figure 3). The global pattern seems similar for all parties: after the right-wing primaries, the number of publication decreases until January 2017, where it starts increasing again to reach its maximum just before the first round, except for FN where the majority of the activity was focused between the selection of their candidate, Marine Le Pen, during the first round, and the second round. Indeed this week represents 26% of FN's tweets, compared to 15% for FI, 9% for PS, 12% for EM, and 10% for LR. This difference is even more visible for retweets, with 27% of FN's retweets published during this week, compared to 12% for FI, 7% for PS, 12% for EM, and 10% for LR.

Profiles seniority. In order to check if the profiles in our dataset were already active or created specifically for the presidential campaign, we looked at their month of creation.

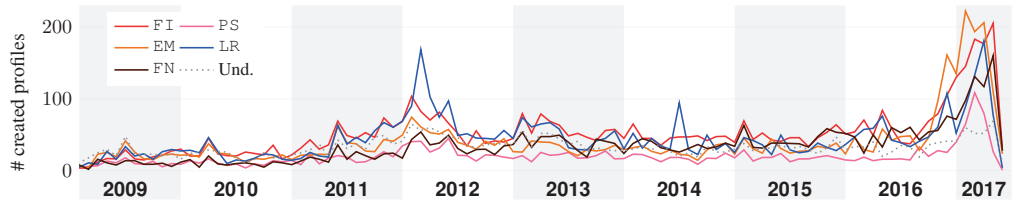

Figure 4: Number of created profiles by month – the 209 profiles created before 2009 do not appear on this graph.

	FI	PS	EM	LR	FN	Rest
FR	2,689	1,090	2,704	2,848	1,723	1,899
US	26	8	35	21	69	35
GB	30	6	40	18	22	25
BE	23	1	15	19	14	27
CA	16	4	15	10	13	13
ES	12	1	13	6	13	25
CH	11	4	10	10	6	10
IT	2	1	10	6	13	13
MA	6	4	4	1	2	23
DE	8	1	6	3	3	8
GR	2	1	2	1	1	15
Other	61	17	57	49	53	117
Total	2,886	1,138	2,911	2,992	1,932	2,210
Abroad	**7%**	**4%**	**7%**	**5%**	**11%**	**14%**

Table 4: Countries present in the dataset by party – only the 5 main parties and the countries indicated by 20 profiles or more are detailed.

For simplicity concerns, only the six main stances are displayed. Results are presented in Figure 4. A large proportion of profiles seems to have been created just before or during the campaign: 19% of the profiles were created after September 2016. This suggests that a majority of these profiles were dedicated to political expression. Apart from these recent creations, the only notable phenomenon is a peak in early 2012, probably connected to the previous presidential election whose rounds happened April 22th April and May 6th 2012. Kolmogorov-Smirnoff and Mann-Whitney tests indicate no significant difference between parties ($\alpha = 5\%$).

5.3 Geographic repartition and languages

This dataset does not exclude non-French tweets to take into account expatriates, French citizens tweeting in another language, and foreigners discussing or participating in the election. Table 4 presents the most represented countries in the dataset, based on the location indicated on the Twitter profiles. We discarded the 192 profiles whose location changed during the campaign, then manually inferred countries when possible, hence determining countries for 86% of the 16,396 annotated profiles whose location was filled in. Table 5 presents the most used languages in the dataset and their use by profiles, as detected by Twitter. The first and secondary languages are determined by the number of published tweets. For this analysis, we did not consider retweets and tweets whose language was undetermined, and we also

	Secondary					
First	French	English	Spanish	Italian	Greek	Total
French	9,376	3,612	506	204	7	**13,705**
English	161	264	7	4	1	**437**
Spanish	12	5	46	1		**64**
Italian	4	5		18		**27**
Greek	3	1			9	**13**
Total	**9,556**	**3,887**	**559**	**227**	**17**	**14,266**

Table 5: First and secondary languages used by profiles in the dataset – only the 5 most used languages are displayed.

Main language	FI	PS	EM	LR	FN	Total
English	50	9	94	10	161	**324**
Spanish	8	1	7	5	17	**38**
Italian	1		3	1	19	**24**
German	1		2		5	**8**
Portuguese	1		1	1	5	**8**
Greek	3		1		1	**5**
Dutch			1		4	**5**
Polish				1	4	**5**
Other	1	2	2	1	7	**28**
Total	**66**	**13**	**112**	**19**	**223**	**433**

Table 6: Parties of profiles whose first language is not French – only languages used by 5 profiles or more are detailed.

discarded tweets classified as being written in Indonesian, Haitian, or Tagalog, since a manual study of these tweets revealed that they were simply misclassified, probably due to their short size and to the presence of a high number of mentions, hashtags, and urls. Despite this election being a French event, it was also discussed abroad, particularly in the neighboring European countries. Interestingly, the proportion of profiles indicating a country other than France is almost double for FN compared to other parties, with 11% of FN's profiles not being located in France. This observation is confirmed by a closer look at the parties associated with the profiles whose first language is not French, as presented in Table 6. The number of non French-speaking FN profiles largely surpasses other parties. Moreover, while for most parties the only notable language is English, FN's profiles are a lot more diverse, with a notable amount of Spanish and Italian profiles.

		Number of elements		Min	Med	Avg	Max	Density	Diameter	Assortativity on party
Retweets	Nodes	22,048	Out-degree	1	14	86	7,183	0.003	333	0.17
			In-degree	1	16	61	1,783			
	Edges	1,321,948	Weight	1	1	3	7,983			
Mentions	Nodes	22,569	Out-degree	1	26	84	2,168	0.004	122	0.14
			In-degree	1	16	107	14,469			
	Edges	1,896,262	Weight	1	1	4	10,473			

Table 7: Retweet and mention networks characteristics.

Figure 5: Retweet network – the size of the nodes is proportional to their out-degree, and the color scheme is identical to Figure 1, with intermediate colors for profiles being affiliated to several parties.

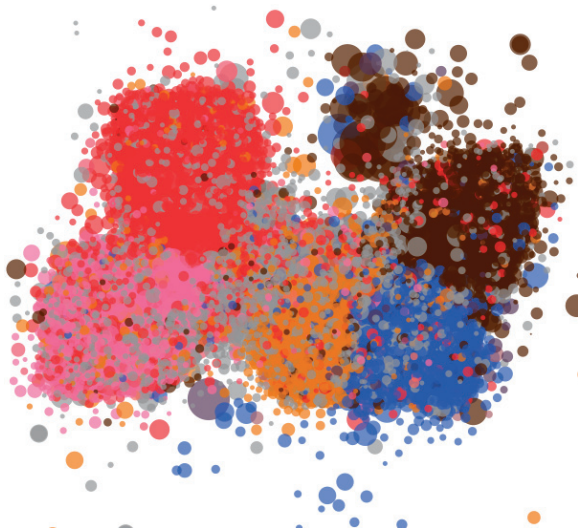

Figure 6: Mention network – the size of the nodes is proportional to their out-degree, and the color scheme is identical to Figure 1, with intermediate colors for profiles being affiliated to several parties.

5.4 Retweet and mention networks

In addition to the profile annotations, this dataset also contains the retweet and mention networks between annotated profiles. This represents 3,203,187 retweets and 6,371,852 mentions, modeled as directed networks following the information flow: from retweeted profiles to retweeting ones, and from mentioning profiles to mentioned ones. The edges' weight represents the number of interactions between the source and the target. Table 7 summarizes their main characteristics, and Figure 5 and Figure 6 offer visualizations, using an OpenOrd layout. The median profile in the dataset tends to retweet a little more than it is being retweeted, and mentions other profiles significantly more than it is mentioned by others. The mention network is denser than the retweet one and is less segregated by political community, as shown by the visual representation and the smaller assortativity – measuring the preference for nodes to attach to similar ones. If we compare with assortativity values given in (Newman 2002), retweet and mention networks are positioned between biology coauthorship and film actor collaborations, denoting a notable preferential attachment for nodes belonging to the same political party.

Interactions between parties. Observing the interactions between parties can be a good way to study the dynamics between them. Table 8 and Table 9 present respectively the mean number of retweets and of mentions aggregated by party. For simplicity concern we only consider the 5 main stances. The results are consistent with the representations in Figure 5 and 6. Unsurprisingly, the overwhelming majority of interactions stays inside the political community. Retweets suggest a proximity between LR and FN, as well as between PS and EM, although it is less pronounced. Men-

		Retweeting party				
		FI	PS	EM	LR	FN
Retweeted party	FI	143	2	1	1	1
	PS	2	82	4		
	EM	3	9	122	3	1
	LR	3	1	3	204	20
	FN	1	1	1	18	194

Table 8: Mean number of retweets by profile according to the retweeted and retweeting parties.

Mentioning party	Mentioned party: FI	PS	EM	LR	FN
FI	249	24	16	11	6
PS	14	240	21	7	3
EM	8	13	232	22	7
LR	4	4	23	382	29
FN	6	4	16	39	301

Table 9: Mean number of mentions by profile according to the mentioning and mentioned parties.

tions are a little less party-based. Almost every party occasionally mention EM and LR profiles, while FI profiles are mainly mentioned by PS ones, FN profiles by LR ones, and PS profiles by FI or EM ones.

6 Limitations

6.1 Twitter limits.

While Twitter is an excellent resource to study political discourse on social media, it is important to stress that one has to be extremely cautious when drawing conclusions on a more general scale (Barberá and Rivero 2015). As summarized by Gayo-Avello (2012), there are three main biases with Twitter data:

- There is no guarantee of veracity – manual annotations partially limit this issue since we can discard obvious bots or parodic profiles, but we cannot vouch for the content of each profile or post;
- The majority of profiles do not tweet about politics: Colleoni, Rozza, and Arvidsson (2014) estimate that around 10% of Twitter discourse is related to political topics. This means that we captured in this dataset a small fraction of Twitter's French user base, which itself represented 5% of the French population in 2013 (IPSOS 2013).
- The demographics of Twitter users vary greatly from France's demographics: Twitter indicates that half of its French users are between 25 and 39, with a 54% / 46% repartition between men and women, and 34% of upper class (Twitter Marketing FR 2016). For comparison, men represents 49% of France's general population, 18% of its population is aged between 25 and 39, and only 15% of the population has an upper class job (INSEE 2010; 2017).

Moreover, the presidential campaign being heavily commented on Twitter, DMI-TCAT periodically reached Twitter's API rate limits. This dataset might then be missing a share of the tweets published by the annotated profiles during the campaign.

6.2 Engagement in the campaign.

This dataset provides information about the profiles' political preferences but without measuring their engagement in the campaign. The political communities shared here are highly heterogeneous in terms of levels of implication and beliefs, gathering professional politicians, activists, enthusiasts, and simple electors alike.

6.3 Networks sampling.

The retweet and mention networks published along this paper are subsets based on the source and target nodes of the interactions being part of the annotated profiles. They represent a sample of the complete retweet and mention graphs of the profiles having interacted during the campaign.

7 Possible use cases

The aforementioned analyses and limitations are presented in order to get an idea of the many possibilities offered by the #Élysée2017fr dataset. We propose some ideas on possible investigations or experimentations which could be attempted:

- Political marketisation and individualisation is a hot topic since its popularisation by Obama's 2012 presidential campaign to identify specific segments of the voting population. During this campaign, it was used by EM to select 6 millions homes to call before the first round of the election. This dataset could be used to understand on which criteria these selections of political communities operate. The main stances are fairly balanced, simplifying the development of N-class stance detection models, be it network-based models thanks to the retweet and mention interactions, or text-based models capitalizing on the tweets content. Moreover, the number of annotations enables the setup of robust supervised models.
- A difficult problem is the identification of swing voters. The presence in #Élysée2017fr of 803 profiles belonging to overlapping political communities could enable researchers to shed light on this problematic, by studying their discourse, connections, and positions in the global network.
- The 2,414,584 tweets provide an excellent ground for opinion and argument mining: during the campaign, profiles used Twitter to support their candidate but also to attack others. Finding these supporting or attacking elements and their targets could provide an interesting map of the presidential campaign, showing the "defense" and "attack" strategies of the different parties.
- These tweets can also be used to improve natural language processing tools for French language on Twitter, since French tweets represent 93% of the corpus.
- A network analysis of each party community could determine which structures are effective in terms of political communication, and improve the detection of "influencers" nodes.
- This campaign was also marked by the propagation of malicious rumours and fake documents, making it a perfect ground to study the mechanisms of smear campaigns propagation. It would be interesting to study this propagation from a structural point of view, in order to see how the rumours moved through the network and which were the most important nodes, but also from a temporal point of view to see the influence of external events on the reception of these rumours by profiles.

This is not an exhaustive list of course, and should not limit research ideas emerging from this data. If need be, this dataset can be further enriched by collecting supplementary information with Twitter API, like the profiles' friends and followers for example.

8 Conclusion

We propose in this work an original, large and complex dataset of 22,853 Twitter profiles engaged in the 2017 French presidential election, annotated by experts, and their corresponding 2,414,584 tweets and 7,763,931 retweets. The profiles are affiliated to several parties, among the 5 main parties which emerged during the campaign, or have undetermined political preferences. We also provide information on the nature of the profiles (individual or collective) and the sex of the profiles' owners. In addition to the several considered stances, it is to our knowledge the first dataset with a large number of profiles, and the first one proposing overlapping political communities, enabling the setup of finer and more ambitious experimentations, such as N-class stance classification or swing-voters identification.

The first analyses show that the use of Twitter varies widely according to the party, with a majority of male participants, and many profiles corresponding to activist groups instead of individuals. A large part of the profiles seem to have been created specifically for the campaign, while another important part was presumably created for the previous presidential election. Despite the election being a national event, it attracted international attention, particularly from France's European neighbours, as demonstrated by the many countries and languages present in the dataset. The retweet and mention networks are highly segregated between parties, as confirmed by the slim numbers of inter-party interactions. Further analyses are needed to fully understand the campaign mechanisms, but this dataset is a valuable base for studying Twitter political discourse or evaluating automatic tools for stance detection or network analysis.

Acknowledgments

This project is co-funded by the European Union – Europe is committed to Midi-Pyrénées with the European fund for regional development. The data was collected and in part annotated by the LisTIC project members ("Early Career Researchers" Project of the French National Research Agency, ref. ANR-16-CE26-0014-01): Julien Figeac (coord.), Xavier Milliner, Pierre Ratinaud, Tristan Salord, Fanny Seffusatti, and Nikos Smyrnaios. Our thanks also go to the rest of the annotators: Hubert Dubois, Gilles Hubert, Hervé Le Borgne, Paul Mousset, Gia-Hung Nguyen, Karen Pinel-Sauvagnat, Thibaut Thonet, and Ronan Tournier.

A Appendix: Structure of the dataset

#Élysée2017fr dataset is available at the following address: https://dataverse.mpi-sws.org/dataverse/icwsm18. The manual annotations shared in `profiles_annotations.csv` are detailed in Table 10. The files `posts_ids_*` gathers tweets and retweets ids, divided according to their author's parties for more flexibility. These posts represent a relational database of 12Go once gathered with Twitter API. The files `networks_*` contains the retweet and mention networks described in Section 5.4, in NCOL and GraphML formats. More details are available in the `README` file.

References

Artstein, R., and Poesio, M. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4):555–596. DOI: 10.1162/coli.07-034-R2.

Barberá, P., and Rivero, G. 2015. Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review* 33(6):712–729. DOI: 10.1177/0894439314558836.

Bhowmick, P. K.; Mitra, P.; and Basu, A. 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *HumanJudge*.

Borra, E., and Rieder, B. 2014. Programmed method. *AJIM* 66(3):262–278. DOI: 10.1108/AJIM-09-2013-0094.

Brigadir, I.; Greene, D.; and Cunningham, P. 2015. Analyzing Discourse Communities with Distributional Semantic Models. In *WebSci*, 1–10. DOI: 10.1145/2786451.2786470.

Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo Chamber or Public Sphere? *Journal of Communication* 64(2):317–332. DOI: 10.1111/jcom.12084.

Conover, M. D.; Ratkiewicz, J.; Francisco, M.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2011a. Political Polarization on Twitter. In *ICWSM*, 89–96.

Conover, M. D.; Goncalves, B.; Ratkiewicz, J.; Flammini, A.; and Menczer, F. 2011b. Predicting the Political Alignment of Twitter Users. In *SocialCom*, 192–199. DOI: 10.1109/PASSAT/SocialCom.2011.34.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382. DOI: 10.1037/h0031619.

Fraisier, O.; Cabanac, G.; Pitarch, Y.; Besancon, R.; and Boughanem, M. 2017. Uncovering Like-minded Political Communities on Twitter. In *ICTIR*. DOI: 10.1145/3121050.3121091.

Garimella, K.; Gionis, A.; Parotsidis, N.; and Tatti, N. 2017. Balancing information exposure in social networks. In *NIPS*. 4666–4674.

Gayo-Avello, D. 2012. No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing* 16(6):91–94. DOI: 10.1109/MIC.2012.137.

Golder, S., and Macy, M. 2015. Opportunities and Challenges for Online Social Research. In *Twitter: A Digital Socioscope*. 1–20.

INSEE. 2010. Des spécificités socioprofessionnelles régionales (fr). www.insee.fr/fr/statistiques/1281090.

INSEE. 2017. Population totale par sexe et âge au 1er janvier 2017, France métropolitaine (fr). www.insee.fr/fr/statistiques/1892088.

Column	Content
FROM_USER_ID	The profile's id used by Twitter
PROFILE_NATURE	"individual" if managed by a single person, else "non individual". The "non individual" aspect can be "political" for parties or groupes of militants, "media" for media outlets, and "other".
PARTY	The profile's political affiliation(s) (see Section 4.2), separated by a slash (ex: "ps/fi").
MEDIA_PROFESSIONAL	Indicates self-identification as a media professional (for individual profiles only).
SEX	Indicates the sex of the owner: "m", "f", or null (for individual profiles only).

Table 10: Content of `profiles_annotations.csv` for each profile.

IPSOS. 2013. Usages et pratiques de Twitter en France. www.ipsos.com/fr-fr/usages-et-pratiques-de-twitter-en-france.

Jungherr, A.; Jürgens, P.; and Schoen, H. 2012. Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions. *Social Science Computer Review* 30(2):229–234. DOI: 10.1177/0894439311404119.

Jungherr, A. 2013. Tweets and votes, a special relationship. In *PLEAD*, 5–14. DOI: 10.1145/2508436.2508437.

Kratzke, N. 2017. The #BTW17 Twitter Dataset–Recorded Tweets of the Federal Election Campaigns of 2017 for the 19th German Bundestag. *Data* 2(4):34. DOI: 10.3390/data2040034.

Lu, H.; Caverlee, J.; and Niu, W. 2015. BiasWatch. In *CIKM*, 213–222. DOI: 10.1145/2806416.2806573.

Makazhanov, A., and Rafiei, D. 2013. Predicting political preference of Twitter users. In *ASONAM*, 298–305. DOI: 10.1145/2492517.2492527.

Newman, M. E. J. 2002. Assortative Mixing in Networks. *Physical Review Letters* 89(20). DOI: 10.1103/PhysRevLett.89.208701.

Pennacchiotti, M., and Popescu, A.-M. 2011. Democrats, republicans and starbucks afficionados. In *KDD*, 430. DOI: 10.1145/2020408.2020477.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond Binary Labels. In *ACL*, 729–740. DOI: 10.18653/v1/P17-1068.

Ratkiewicz, J.; Conover, M.; Meiss, M.; Gonçalves, B.; Patil, S.; Flammini, A.; and Menczer, F. 2011. Truthy. In *WWW*, 249–252. DOI: 10.1145/1963192.1963301.

Saez-Trumper, D. 2014. Fake tweet buster. In *HT*, 316–317. DOI: 10.1145/2631775.2631786.

Sang, E. T. K., and Bos, J. 2012. Predicting the 2011 Dutch Senate Election Results with Twitter. In *Workshop on Semantic Analysis in Social Media*, 53–60.

Sunstein, C. R. 2009. *Republic. Com 2. 0.* ISBN: 978-0-691-14328-6 978-1-4008-2783-1 978-1-282-75459-1.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *ICWSM*.

Twitter Marketing FR. 2016. Utilisateurs de Twitter en France. https://twitter.com/TwitterMktgFR/status/781038037669216256.

Vosoughi, S.; Mohsenvand, M. N.; and Roy, D. 2017. Rumor Gauge. *TKDD* 11(4):1–36. DOI: 10.1145/3070644.

Weber, I.; Garimella, V. R. K.; and Batayneh, A. 2013. Secular vs. Islamist polarization in Egypt on Twitter. In *ASONAM*, 290–297. DOI: 10.1145/2492517.2492557.

Weber, I.; Garimella, V. R. K.; and Teka, A. 2013. Political Hashtag Trends. In *Advances in Information Retrieval*, volume 7814. 857–860. DOI: 10.1007/978-3-642-36973-5_102.