“ L’avenir en commun ” des Insoumis. Analyse des forums de discussion des militants de la France Insoumise

Clément Plancq, Zakarya Després, Julien Longhi

« L'avenir en commun » des Insoumis. Analyse des forums de discussion des militants de la France Insoumise

Clément Plancq*, Zakarya Després**
Julien Longhi***

*Lattice, ENS, CNRS, Univ. Paris 3, 1, rue Maurice Arnoux. 92120 Montrouge
clement.plancq@ens.fr,
http://www.lattice.cnrs.fr/plancq
**Lattice, ENS, CNRS, Univ. Paris 3, 1, rue Maurice Arnoux. 92120 Montrouge
zakarya.despres@gmail.com
***Université de Cergy-Pontoise. Laboratoire AGORA
33 Boulevard du port. 95011 Cergy-Pontoise
julien.longhi@u-cergy.fr
https://www.u-cergy.fr/fr/_plugins/mypage/mypage/content/jlonghi.html

Résumé. Les discours politiques ont fait l'objet de travaux marquants en analyse du discours et en TAL mais les études sur les discussions de militants sont plus rares. Pourtant ces communautés sont le lieu d'échanges idéologiques sur le programme d'un candidat. L'étude de ces discussions peut se révéler intéressante pour étudier la circulation des idéologies de l'appareil politique vers une communauté de citoyens et vice-versa.

Dans l'article nous présentons les travaux menés pour recueillir un corpus de messages émanant de forums de discussion des militants de la France Insoumise puis les analyses conduites sur ce corpus à l'aide des outils de la plateforme Cortext.

1 Introduction

Les travaux que nous présentons ont été menés pendant le datasprint datapol [1]. Plus précisément ce travail s'inscrit dans le projet « #Présidentielles 2017 : comparaison et circulation des idéologies ». Parmi les questions abordées dans ce projet, nous nous sommes intéressés à la porosité entre les idéologies des candidats et les communautés des militants.
Un datasprint est un exercice stimulant avec des contraintes fortes. L'étude présentée en est empreinte : d'une part, après seulement 3 jours de travail sur les données notre étude est nécessairement exploratoire et d'autre part nous nous sommes concentrés sur les idéologies, laissant de côté une part non négligeable du contenu du corpus. Pour ces recherches, le Lattice a bénéficié d'un financement de PSL (Paris Sciences et Lettres, ref. ANR-10-IDEX-0001-02 PSL), dans le cadre de l'appel à projets SHS 2016.

1. datapol (http://bit.ly/data-pol) organisé par le medialab de Sciences-po, du 29 novembre au 2 décembre 2017.

Plusieurs travaux, chercheurs, ou projets, ont analysé les discours politiques lors des récentes campagnes électorales. Le projet « Mesure du discours » piloté à Nice par Damon Mayaffre a contribué à rendre accessible l'analyse des discours politiques de la campagne présidentielle 2017, dans Mayaffre et al. (2017) notamment. Les travaux de Alduy (2017) ont proposé une analyse de 1300 textes (2,5 millions de mots - écrits ou prononcés de 2014 à 2016) des principaux candidats avec le logiciel Hyperbase. D'autres travaux se sont intéressés aux discours politiques numériques, tels que le projet #Idéo2017 piloté par Julien Longhi à l'université de Cergy-Pontoise. Ce projet s'appuie notamment sur une caractérisation préalable du tweet politique comme genre de discours (Longhi (2013)), et confère une légitimité aux analyses de discours natifs du web.
À notre connaissance les forums des militants n'ont pas encore fait l'objet de telles attentions. Pourtant les forums de discussion sont une source de données souvent exploitée. Au point d'avoir suscité des reflexions méthodologiques ; ainsi pour Marcoccia (2004) [2] « il s'agit d'un corpus idéal pour l'analyse des conversations et l'analyse du discours, car il répond aux critères suivants : 1. Il s'agit d'échanges authentiques produits en l'absence de l'analyste qui les enregistre, ce qui permet d'éviter un des problèmes méthodologiques habituels de l'analyse des conversations [...] 2. Ces corpus sont homogènes, définis par leur mise en mémoire [...] ». Pour les militants et les sympathisants d'un homme ou d'un parti politique, les réseaux sociaux numériques sont des supports d'échanges et de dialogues privilégiés que le chercheur peut exploiter. Notre objectif ici est de mesurer la diffusion de l'idéologie de la France Insoumise en cherchant, dans l'analyse des productions de la base militante, les traces des idéologies saisissables sous formes de doxas. Pour Longhi et Sarfati (2012) la doxa, par ses contenus comme par ses formes expressives, « tend à verser dans le domaine public, au-delà de l'institution de sens dont elle tire ses contenus minimaux. Elle se distingue par son haut degré de stéréotypie, ainsi que par une hétérogénéité non marquée. Enfin, la doxa est le lieu par excellence de la naturalisation d'un discours ».

Nous nous sommes focalisés sur le candidat Jean-Luc Mélenchon : sur son programme « L'avenir en commun » (LAEC) et sur la communauté militante de la France Insoumise (les Insoumis) telle qu'elle a pu s'exprimer dans les forums de discussion numériques. Sur le plan numérique le candidat Mélenchon se distingue des autres candidats aux élections présidentielles de 2017. D'abord parce qu'il a placé le numérique au centre de sa stratégie de campagne : en plus d'être présent sur les réseaux sociaux, il fut le premier en France à utiliser le logiciel NationBuilder et à avoir une chaîne Youtube. Mais surtout le candidat a été soutenu par des relais militants auto-organisés avec le « Discord insoumis » dont l'idée a germé dans un forum de discussion de jeuxvideo.com. Ces deux espaces numériques distincts, l'un officiel, l'autre participatif, se prêtent bien à notre objectif d'analyse de la circulation des idéologies.

Nous nous attarderons dans un premier temps sur les étapes de constitution d'un corpus de travail issu des forums de discussion des Insoumis. Puis nous détaillerons deux analyses outillées de ce corpus.

2. l'article porte sur les forums `usenet`.

2 Les données

2.1 Recrutement des données

Pour rassembler les conversations des Insoumis, nous nous sommes intéressés à deux espaces de discussion : le forum Blabla 18-25 de jeuxvideo.com et le Discord des Insoumis.
Si comme son nom l'indique, jeuxvideo.com est à la base un site consacré aux actualités sur les jeux vidéo, au sein de son forum ce sont les différentes sections de discussions générales qui y sont les plus populaires [3]. On y trouve notamment le "Blabla 18-25", où se retrouvaient de nombreux Insoumis pendant la campagne. Ils se sont notamment rassemblés sur une série de sujets, spécifiquement dédiés à Jean-Luc Mélenchon et France Insoumise.
Depuis 2016, le site jeuxvideo.com ne propose plus d'API publique pour accéder à ses données, nous avons donc écrit un script de web scraping en Python, grâce à la bibliothèque logicielle BeautifulSoup [4]. Nous avons ainsi pu récupérer 21 4761 messages, postés sur une série de sujets consacrés à Jean-Luc Mélenchon et la France Insoumise sur une période s'étendant du 12/11/2016 au 31/10/2017.
C'est aussi du forum Blabla 18-25 qu'est né le Discord des Insoumis, second espace de discussion que nous avons étudié. Discord est un logiciel gratuit de VoIP, conçu à la base pour les joueurs de jeux vidéo. Son avantage, par rapport à ses alternatives comme Slack ou Skype, est de permettre de converser dans des salons vocaux et des salons textuels en parallèle. Pour une communauté comme celle des Insoumis, c'est un moyen d'organiser des débats sous plusieurs formes, à l'écrit comme à l'oral, voire même de produire des podcasts comme ceux de Radio Insoumise.
Contrairement à jeuxvideo.com, Discord possède une API, permettant de récupérer et de poster du contenu. Avant de demander des données d'un serveur, il faut d'abord avoir un compte utilisateur qui y ait accès, afin d'obtenir un token d'authentification qui va nous autoriser à faire des requêtes via l'API. Ces requêtes HTTP retournent du contenu au format JSON dans lequel on trouve le message, sa date mais aussi les liens externes vers des articles ou des images qui ont été partagés par les utilisateurs. Le corpus basé sur Discord est donc constitué de 509 765 messages, datés du 07/02/2017 au 30/10/2017.

2.2 Préparation des données

La préparation des données a été effectuée en amont du *datasprint* afin de les rendre disponibles aux participants de l'événement. Les données étant sous Copyright il n'était pas question de les rendre publiques mais même sous couvert d'un accès restreint il a fallu les modifier afin d'assurer d'une part l'anonymat des messages et d'autre part la facilité d'utilisation des données. Les six fichiers JSON récoltés dans la phase de collecte ont été transformés en fichiers tabulaires comportant les colonnes suivantes :

— `date` : date du message au format Y-m-d ;
— `time` : heure du message au format H-M-S ;
— `content` : le contenu du message ;
— `attach` : l'URL des fichiers attachés aux messages discord ;
— `embed` : l'URL des fichiers inclus dans les messages discord

3. source : http://jvstats.forum-stats.org/stats/1/
4. https://www.crummy.com/software/BeautifulSoup/

Pour les besoins de l'analyse les six fichiers tabulaires ont été agrégés dans une structure de données (*dataframe*) de la bibliothèque pandas [5]. Les données ont été indexées sur le champ `date`, ce qui a facilité les extractions et les analyses situées sur des intervalles de temps choisis.

	messages	mots	dates	origine
jvc.csv	214 761	6 760 746	12/11/2016 - 30/10/2017	jeuxvideo.com
blabla.csv	47 790	631 751	03/09/2017 - 30/10/2017	discord
debat_actu.csv	4 820	88 045	15/10/2017 - 30/10/2017	discord
debat_direct_an.csv	15 518	201 044	10/07/2017 - 30/10/2017	discord
discussion_fi.csv	430 475	6 670 480	07/02/2017 - 30/10/2017	discord
radio_insoumise.csv	11 162	145 589	09/05/2017 - 30/10/2017	discord
total	**724 526**	**14 497 655**		

TAB. 1 – *Détail du corpus France Insoumise.*

Le recueil des discussions des Insoumis présenté dans tableau 1 n'a pas prétention à constituer un corpus exhaustif ou même représentatif. Le corpus de la France Insoumise (corpus FI) ne contient pas les messages antérieurs au 12/11/2016 postés sur le forum Blabla 18-25, il n'inclut ni les dicussions orales tenues dans le discord insoumis ni les canaux de discussion fermés ou archivés entre-temps. Le corpus FI nous semble néanmoins suffisamment volumineux et échelonné dans le temps pour permettre une analyse outillée.

3 Analyses et résultats

Les analyses que nous présentons [6] ont été menées à l'aide des services de la plateforme CorTexT [7] de l'Ifris [8]. L'intégralité de notre corpus a pu y être prise en charge ; chaque analyse a bénéficié du même *modus operandi* inspiré de Chavalarias et Cointet (2008) : extraction des termes [9], tri et nettoyage des termes proposés, indexation des messages avec les termes et enfin cartographie des cooccurrences des termes dans les messages à l'aide du script CorTexT « network mapping ».

3.1 De quoi discutent les militants ?

La première analyse a consisté à identifier et cartographier les thématiques abordées dans les discussions des militants de la France Insoumise. Nous avons extrait une liste de 500 termes candidats de l'ensemble du corpus en employant les paramètres suivants : chi-2 pour le score de spécificité, occurrence minimale de 3, longueur maximale de 3 tokens. Cette liste a ensuite été nettoyée manuellement, nous l'avons expurgé de termes non pertinents (*seul truc*, *mais bon*, *bonne nuit*, etc . . .). Puis l'ensemble des messages a été indexé en fonction de la liste de

5. https://pandas.pydata.org/

6. En plus des auteurs, trois étudiants du Master de sociologie des mondes numériques de Marne-La-Vallée ont participé à ces analyses. Emmanuelle Coniquet, Amalia Nikolaidi et Adil Ouafssou.

7. http://www.cortext.org/

8. http://ifris.org/

9. l'extraction de CorTexT porte sur les formes racinisées

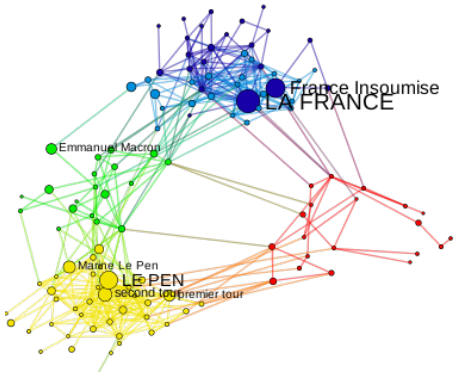

FIG. 1 – *Réseau des termes clustérisés sur l'ensemble du corpus.*

termes.
La figure 1 est le résultat de l'analyse en réseau de CorTexT, elle représente la carte des cooccurrences des termes indexés dans les messages du corpus. Dans la figure chaque nœud est un terme associé à sa valeur de cooccurrence, nous avons restreint le nombre de noeuds du graphe à 150. Les arêtes entre les noeuds sont pondérées par une mesure de proximité (ici la similarité cosinus). Les grappes de points du graphe forment des communautés ou clusters, l'algorithme de détection de communautés utilisé est celui de Louvain. Le partitionnement en sous-graphes y est basé sur la maximisation de la modularité de Newman (2006).
Dans la figure, plus deux termes sont proches plus ils sont associés fréquemment dans le corpus. Plus un terme apparaît conjointement avec d'autres termes, plus le point est gros.
Les clusters de couleurs jaune et vert clair de la figure 1 regroupent les discussions sur les autres candidats (Le Pen, Macron, Fillon, Hamon), le cluster rouge rassemble les préoccupations d'organisation interne du groupe et les actions de soutien au candidat Mélenchon, le cluster bleu réunit les discussions idéologiques sur le programme « L'avenir en commun ».
La figure 2 offre un zoom sur le sous-graphe idéologique et permet d'y distinguer des termes saillants comme : *fraude fiscale*, *transition énergétique*, *service public*, *traités européens*, *assemblée constituante*, etc … autant de mots clés du programme du candidat Mélenchon, tous reliés aux deux points centraux du sous-graphe que sont *LA FRANCE* et *France insoumise*. Cette analyse montre clairement qu'une partie importante des discussions porte sur l'idéologie.

3.2 « L'avenir en commun » vus par les Insoumis

En second lieu nous avons cherché à confronter les messages du corpus avec le texte du programme LAEC. Ce programme est une production du parti politique la France Insoumise, il a d'abord été publié aux éditions du Seuil avant que les Insoumis n'en fassent le site http://laec.fr [10]. Il s'agit d'un texte qui fait œuvre de communication politique, son contenu

10. cette initiative de publication électronique est d'ailleurs discutée sur le discord insoumis et se retrouve dans notre corpus.

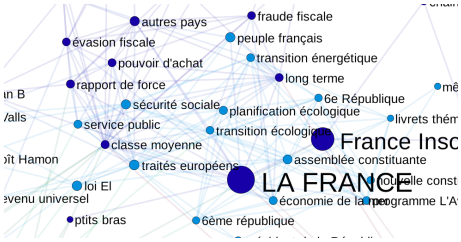

FIG. 2 – *Zoom sur les termes du cluster idéologique.*

est choisi, la langue est soigneusement vérifiée alors que notre corpus recèle des productions écrites spontanées où la syntaxe et l'orthographe ne sont pas toujours corrigées. De plus les messages n'ont pas forcément une sémantique autonome, ils peuvent faire partie d'un fil de discussion et faire référence de manière elliptique à des arguments exposés antérieurement. Ainsi par exemple : « pour recentrer un peu sur ce que tu disais, je pense sincèrement qu'il n'y a pas de vrai vote utile. » (discord, 04/04/2017).
LAEC et le corpus FI sont donc très différents par leur forme. Par leur volume également, nous n'avons pas pu compter la taille en nombre de mots du programme mais les 128 pages du livre forment un ensemble qui n'est pas statistiquement comparable avec les 13 millions de mots du corpus FI.

L'objectif de cette seconde analyse est de découvrir en quels termes sont discutés chacun des 7 chapitres du programme LAEC dans le corpus. Pour cela nous avons extrait manuellement 150 termes du programme. Puis la sous-partie du corpus antérieure au 23 avril 2017, date du premier tour de l'élection présidentielle, a été indexée en fonction de la liste de termes. À nouveau CorTexT a été sollicité pour produire un graphe de cooccurrences. Nous avons utilisé la même méthode qu'en 3.1.
6 clusters différents se distinguent dans la carte générée. Nous nous sommes intéressés cette fois aux labels des clusters. Ceux-ci sont proposés automatiquement, les labels étant les termes les plus centraux d'un cluster donné, c'est-à-dire ceux qui ont le plus de relations avec l'ensemble des termes du cluster.

chapitres	labels
La 6ème République	nouvelle constitution & constituante
Protéger et partager	smic & dette
La planification écologique	international & climat
Sortir des traités européens	UE & Europe
Pour l'indépendance de la France	humanisme & privatisation
Le progrès humain d'abord	militant & mouvement
La France aux frontières de l'humanité	

TAB. 2 – *Titres des chapitres de LAEC et labels.*

Le tableau 2 rapproche les intitulés des chapitres de LAEC avec les labels des clusters de la seconde analyse. Si un label comme *militant & mouvement* appartient plutôt exclusivement à la sphère militante, les quatre labels restants partagent des proximités sémantiques avec les chapitres de LAEC : *nouvelle constitution & constituante* et *La 6ème République*, *UE & Europe* et *Sortir des traités européens* par exemple. Nous pouvons émettre l'hypothèse qu'il s'agit là de reformulations par les militants des thèmes du programme du candidat. Signe d'une véritable porosité entre la communication officielle et les réseaux sociaux auto-organisés des militants.

4 Conclusion

Nous avons présenté les fruits du travail que nous avons pu mener pendant les trois jours du datapol, ils ne constituent qu'une première exploration mais les analyses sont encourageantes. L'analyse outillée a confirmé notre hypothése selon laquelle une part non négligeable des discussions des militants de la France Insoumise portent sur les idéologies défendues par le candidat et son programme.
Ce travail soulève surtout des questions que le format du *datasprint* ne nous a pas permis d'examiner : comment quantifier l'importance des discussions idéologiques comparées à l'ensemble du corpus ? Comment qualifier ces discussions ? Nous savons que les thèmes du programme ont été discutés mais nous ignorons dans quelle mesure ils ont fait l'objet de controverses. La question pourra être traitée dans un travail ultérieur en appliquant des techniques d'analyse de sentiment issues du TAL ou à l'aide d'indices linguistiques de négation ou de polarité négative.

Références

Alduy, C. (2017). *Ce qu'ils disent vraiment. Les politiques pris aux mots*. Le Seuil.

Chavalarias, D. et J.-P. Cointet (2008). Bottom-up scientific field detection for dynamical and hierarchical science mapping - methodology and case study. *Scientometrics 75*(1), 20.

Longhi, J. (2013). Essai de caractérisation du tweet politique. *L'information grammaticale 136*, 25–32.

Longhi, J. et G. Sarfati (2012). *Dictionnaire de pragmatique*. Colin.

Marcoccia, M. (2004). L'analyse conversationnelle des forums de discussion : questionnements méthodologiques. *Les Carnets du Cediscor 8*.

Mayaffre, D., C. Bouzereau, M. Ducoffe, M. Guaresi, F. Precioso, et L. Vanni (2017). Les mots des candidats, de " allons " à " vertu ". In P. Perrineau (Ed.), *Le vote disruptif. Les élections présidentielle et législatives de 2017*, pp. 129–152. Presses SciencesPo.

Newman, M. E. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci U S A 103*(23), 8577–8582.

Summary

Political speeches have been the focus of discourse analysis and NLP, but studies of activist discussions are scarcely found. Yet these communities are the place of ideological exchanges

on the party platform. The study of these discussions can be interesting to study the circulation of ideologies of the political apparatus towards a community of citizens and vice versa.

In the article we present the work carried out to gather a corpus of messages emanating from forums of discussion of the militants of France Insoumise. Then we discuss the analyzes conducted on this corpus using the tools of the platform Cortext.