POLI-PRESS-FR : Un Corpus de Communiqués de Presse des partis politiques français pour l'analyse de discours et l'apprentissage automatique

MADICS, Avignon Juin 2026

Contexte

L'intersection croissante entre la recherche d'information, l'apprentissage automatique et les sciences sociales computationnelles a engendré un besoin important de jeux de données spécialisés et de haute qualité pour étudier les phénomènes sociétaux complexes.

Pour répondre à ce besoin, nous présentons POLI-PRESS-FR, un nouveau corpus longitudinal de communiqués de presse (CPs) officiels de cinq partis politiques français (Rassemblement National, Les Républicains, La République en Marche, Parti Socialiste et La France Insoumise).

La création de cette ressource a nécessité un processus de collecte méticuleux, rendu indispensable par l'absence d'archivage systématique et la volatilité des sites web des partis politiques, permettant ainsi de surmonter d'importantes difficultés d'acquisition de données.

Introduction

Bien que les communiqués de presse puissent sembler proches du genre plus large de l'information politique, ils s'en distinguent fondamentalement :

- Rédigés, validés et publiés par les partis politiques eux-mêmes, les communiqués de presse constituent une preuve directe de leur discours officiel.
- Leur contenu, souvent limité dans sa portée, mais offrant une plus grande liberté thématique que d'autres formats de communication électorale (Gessler et Hunger, 2022), est sélectionné en fonction d'un agenda politique interne, et leur fonction principale est d'introduire et de diffuser des thèmes spécifiques au sein du débat public.

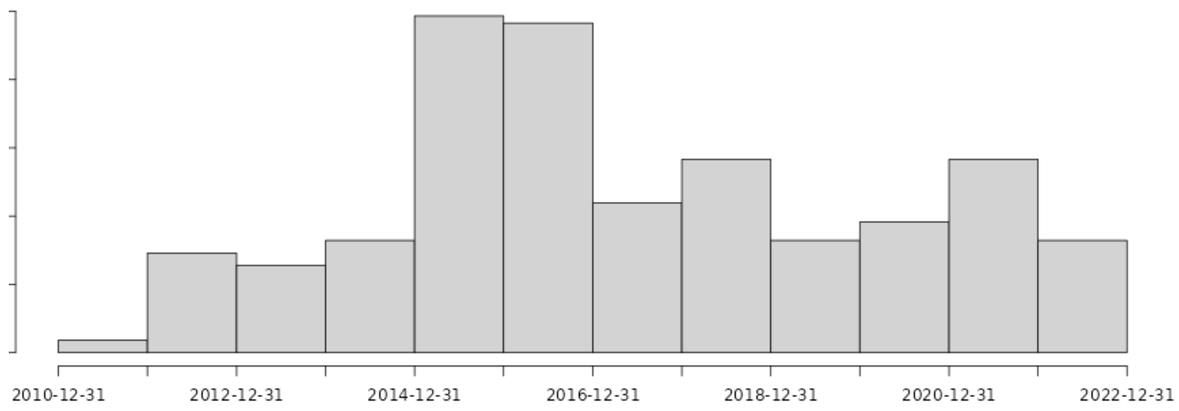
Les pages web des partis extraites étaient les suivantes :

- *RN :* `https://rassemblementnational.fr/communiques`,
- *PS :* `https://www.parti-socialiste.fr/dernier_cp`,
- *LR :* `http://www.republicains.fr/communiques`,
- *LREM :* `https://en-marche.fr/articles/communiques/`,
- *FI :* `https://lafranceinsoumise.fr/category/actualites/communiques/`.
- `https://archive.org/`

Le Corpus

Année \ Parti	FI	LR	LREM	PS	RN	Total
2016	5	274	1	388	731	1399
2017	17	66	129	299	451	962
2018	104	16	48	108	630	906
2019	140	31	32	45	379	627
2020	96	14	19	113	382	624
2021	113	37	20	108	345	623
2022	83	17	0	77	228	405
2023	181	30	0	167	231	609
2024	159	15	0	141	91	406
TOTAL	898	500	249	1446	3468	6561

Nombre de CPs par Parti et par Année



Migrants et frontières, histogram of a pool of 300 documents from RN based on dot product similarity

Méthodologie

Une approche unique est insuffisante pour saisir la complexité du langage politique. Nous proposons une approche hybride qui combine la recherche et le classement denses avec l'extraction stochastique d'ensembles d'éléments fréquents, selon les principes de l'analyse formelle de concepts :

1. Recherche par pool dense flou.
2. Extraction stochastique d'ensembles de mots fréquents.
3. Analyse formelle de concepts (AFC).
4. Analyse de texte mixte dense/clairsemé.
   - Recherche dense
   - Raffinement
   - Hybridation

   Cette approche hybride permet au chercheur de se concentrer sur un groupe spécifique de termes, puis de prendre du recul pour identifier un ensemble de documents pertinents au sein du périmètre thématique initial.

Extreme right Frequent item sets. All data from le rassemblement national. $k = 12$*, cardinal extension set at 5.*

extension	nb
plus & français & marine & pen & politique	525
européen & europe & parlement & européenne & député	450
front & national & immigration & français & nicolas	156
bien & plus & faire & depuis & tout	94
contre & ainsi & communiqué & france & toujours	72
groupe & france & dont & président & rassemblement	60
immigration & politique & asile & france & clandestins	56
comme & aussi & pays & être & donc	45
frontières & européenne & union & plus & schengen	38
plus & français & sécurité & depuis & mettre	27

Quelques statistiques

Statistic \ Party	FI	LR	LREM	PS	RN
Text Mean NoW	264.1	249.0	333.5	289.3	255.3
Text Median NoW	224	232	219	242	242
Text Mean NoC	1685.8	1585.5	2158.9	1845.5	1633.6
Text Median NoC	1415	1485	1414	1545	1552
Title Mean(Median) NoW	12.1(12)	11.6(11)	12.2(13)	12.8(12)	11.9(11)
Text Total Tokens	247044	208565	83285	488706	1636518
(normalized)	(262.81)	(247.11)	(334.48)	(278.64)	(254.67)
Text Vocabulary	19619	18146	10904	24389	50362
(normalized)	(20.87)	(21.5)	(43.79)	(14.35)	(7.84)
Title Vocabulary	2892	2685	803	3723	10527
(normalized)	(3.07)	(3.18)	(3.22)	(2.19)	(1.64)

Statistiques sur les CPs (v2)

Conclusion

Nous présentons ici POLI-PRESS-FR, un nouveau corpus longitudinal de communiqués de presse officiels des cinq principaux partis politiques français (`https://pol.termwatch.eu/`).

Ce corpus a nécessité d'un processus rigoureux de collecte semi-manuel et de curation pour surmonter le défi majeur de la volatilité des données.

Nous utilisons une méthodologie hybride LDA-FCA pour identifier et suivre l'évolution de thèmes politiques spécifiques

Ce corpus peut aussi être utilisé pour évaluer les LLM légers afin d'explorer leurs orientations politiques.

J. Vermeirsche, T. Jiménez et E. SanJuan

Jeanne Vermeirsche est post-doctorante et a fait sa thèse au Laboratoire .JPEG, Tania Jiménez et Eric SanJuan sont au Laboratoire Informatique d'Avignon – LIA UPR 4128