Explorer les mots du politique dans la transformation numérique. Analyser le lexique politique dans des contextes et selon des ressources en évolution

Julien Longhi

Lingue e Linguaggi
Lingue Linguaggi 65 (2024), 361-384
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v65p361
http://siba-ese.unisalento.it, © 2024 Università del Salento

EXPLORER LES MOTS DU POLITIQUE DANS LA TRANSFORMATION NUMERIQUE
Analyser le lexique politique dans des contextes et selon des ressources en évolution

JULIEN LONGHI
CY CERGY PARIS UNIVERSITÉ, AGORA LAB, INSTITUTE OF DIGITAL HUMANITIES

Abstract – The evolution of digital technologies has profoundly transformed the way we exchange political ideas. The rise of social networks, blogs, and online discussion forums has provided an increasingly accessible platform for political organizations to communicate and interact with voters. In order to explore words in digital transformation, and the use of tools and/or approaches for studying lexicon and phraseology in evolving discursive domains, we chose to focus our study on words in the political domain, in the context of the 2017 and 2022 French political elections. Thus, in the interval of two presidential campaigns, the political context has changed enormously (evolution of the French political landscape, recomposition of parties and the electorate), and corpus analysis technologies have also undergone a great evolution. The analysis of digital political discourse has become increasingly important due to the growing importance of the Internet and social media in public debate and opinion formation. New challenges in digital political discourse analysis include the sheer volume of data (which is often noisy and contains redundant or irrelevant information), polarization, misinformation, and the difficulty of distinguishing between facts, opinions, and rumors, especially in short messages. This paper will therefore address both the methodological and technological transformations, as well as the discursive and argumentative transformations, of the analysis of political words in an electoral context, through the presentation of different projects and initiatives that have marked the scientific landscape during the campaigns. Finally, to deepen this inventory, and to address the issue of lexicon and political discourse, we will focus on the theme of the analysis of the candidates' style, by highlighting the way in which the use of deep learning and textual statistics can help to better understand the evolution of political discourse, and to measure the contribution of recent technologies and tools mobilizing Artificial Intelligence.

Keywords: political lexicon; textual statistics; digital corpus; artificial intelligence; style.

Le recours à l'ordinateur, rendu nécessaire à la fois par l'ampleur des opérations de dépouillement envisagées et par le volume des calculs statistiques couramment effectués, implique en retour que l'on définisse de manière toujours plus précise l'ensemble des règles qui présideront au dépouillement automatique des textes stockés en machine.
(L. Lebart & A. Salem « Statistique Textuelle », 1994, p. 18).

1. L'analyste du discours face aux évolutions linguistiques et technologiques : enjeux et perspectives

Afin de rendre compte, de manière linguistique, des domaines discursifs en évolution, nous avons choisi de concentrer notre étude sur les mots du domaine politique, dans le cadre des élections politiques françaises de 2017 et 2022. Ainsi, dans l'intervalle de deux campagnes présidentielles, le contexte politique a énormément changé (évolution du paysage politique français, recomposition des partis et de l'électorat), et les technologies d'analyse de corpus ont également connu une évolution spectaculaire. La littérature scientifique en analyse du discours (notamment) témoigne de ces évolutions, tout comme différents projets de recherche qui ont, lors de ces élections, proposé des analyses et produit des articles ou des ouvrages.

Cet article abordera donc à la fois les transformations méthodologiques, technologiques, mais aussi discursives et argumentatives, de l'analyse des mots politiques en contexte électoral (à la suite de Longhi 2016 et 2020), à travers la présentation de deux projets et initiatives qui ont marqué l'actualité scientifique lors des campagnes. Cela permettra de présenter à la fois les enjeux sociétaux et informationnels de tels projets, mais surtout de décrire précisément les outils et technologies mobilisées, et leur évolution sur la période 2017-2022. Nous présenterons ensuite deux projets menés à CY Cergy Paris université, qui s'intègrent dans le mouvement des deux précédemment décrits : #Idéo2017 et PolitiQuiz. Enfin, pour approfondir cet inventaire, et aborder la problématique du lexique et du discours politique, nous approfondirons le thème de l'analyse du style des candidat-e-s, en mettant en valeur la manière dont l'usage du *deep learning* et des statistiques textuelles peuvent aider à mieux comprendre l'évolution du discours politique, et permettre de mesurer l'apport des récentes technologies et des outils mobilisant l'Intelligence Artificielle.

2. Des projets emblématiques d'analyse du discours politique : objectifs et assise scientifique

2.1. Politoscope et Gargantext : analyse communicationnelle et politique

Le Projet Politoscope, CNRS Institut des Systèmes Complexes Paris Ile-de-France (ISC-PIF), http://politoscope.org aborde des questions politiques d'un point de vue informatique et communicationnel, mais pose des questions tout à fait utiles pour des travaux en analyse du discours : ces chercheurs questionnent en effet la manière dont les différentes communautés politiques s'organisent sur Twitter (devenu depuis X[1]) pour diffuser de l'information, ils recensent les thèmes de prédilection des candidats, et leur appropriation par les communautés, ils observent la création et la propagation des infos/fake news, et cherchent à cartographier l'importance des candidats.

Pour l'analyse textuelle, le Politoscope utilise Gargantext, défini comme un « outil de text-mining » : en contexte électoral, les équipes ont « analysé les mesures proposées par l'ensemble des candidats ainsi que tous leurs tweets captés par le Politoscope » afin de « délimiter un vocabulaire spécifique au discours politique […] dont les termes ont été catégorisés en grands thèmes de manière semi-automatique », même si les auteurs reconnaissent quelques erreurs de classification. L'exemple donné sur le site de Gargantext permettra de clarifier l'approche :

> Pour ne prendre qu'un exemple, sous la catégorie Économie & fiscalité sont regroupés tous les tweets qui mentionnent au moins une fois l'un des termes suivants : entreprise*; pme; pmi; actionnaire*; relocalisation*; crédit impôt; nationalisation*; commerçant*; plus-values; délocalisation*; cotisations sociales; scop; niches fiscales; gafa; cice; taux d'imposition; retenue à la source; crédit impôt-recherche; impôt sur les sociétés; austérité; fisc*; impôt; imposition; cotisations sociales; retenue à la source; etc. Lors de la présidentielle 2017, dans un souci d'équité et afin d'être le plus exhaustif possible, nous avons invité le 28 février 2017 toutes les équipes de campagne à nous communiquer des éléments concernant leur programme et les mot-clés qu'ils souhaitaient mettre en avant.

On comprend donc que des catégories/thèmes de campagne ont été définis, et qu'ils sont caractérisés par un vocabulaire, qui permet ensuite de décider la thématique d'un tweet. Gaumont *et al.* (2018, pp. 10-11) précisent la méthode :

[1] Nous conservons dans cet article le nom Twitter quand il s'agit de la plateforme telle qu'elle était considérée/analysée avant le changement de nom, le 23 juillet 2023. Le terme tweet(s) est conservé, indépendamment de la période, pour désigner les messages produits sur Twitter/X.

> Pour mieux caractériser les communautés politiques, reconstruites uniquement à partir de l'analyse des interactions sociales, nous avons analysé leur profil sémantique. Notre objectif n'est pas de les caractériser de manière exhaustive, mais de donner un aperçu de la façon dont les communautés diffèrent dans leur agenda politique et dans leur manière de contextualiser l'information. Pour ce faire, nous avons effectué une analyse du discours politique afin d'extraire les principaux débats de cette campagne. Nous avons utilisé pour cela deux types de sources : les mesures politiques rapportées dans les programmes de campagne, écrites dans un langage contrôlé, et les tweets des candidats, qui empruntent souvent au langage parlé. […] nous avons analysé ces deux corpus à l'aide du logiciel de fouille de texte du CNRS Gargantext (https://gargantext.org ; open source à https://github.com/ISCPIF/gargantext), qui met en œuvre des méthodes avancées de traitement automatique du langage et de détection de thématiques par l'analyse des co-termes ; ainsi que des méthodes de visualisation pour la représentation des résultats et la navigation interactive. Dans les premières étapes, Gargantext utilise une combinaison de lemmatisation, de post-tagging, d'analyse statistique comme tf-idf [47] et d'analyse des relations de généricité entre termes [48] pour identifier dans le texte quelques milliers de mots clés qui sont spécifiques au discours politique. Ces mots-clés ont ensuite été triés par les 377 auteurs afin de sélectionner les plus significatifs (i.e. les mots vides ou les expressions mal formées qui n'auraient pas été filtrée lors des étapes de text-mining ont été supprimés, des hashtags ou néologismes importants de Twitter comme frexit ont été ajoutés). Enfin, toutes les mesures politiques ont été relues attentivement avec les mots-clés sélectionnés mis en évidence dans le texte afin de vérifier qu'il ne manquait aucun mot-clé important. Cela a conduit à un vocabulaire de près de 1600 groupes de mots-clés qualifiant les thèmes de la campagne présidentielle (voir Texte I dans S1 File pour la liste des mots clés). Nous avons utilisé la mesure de proximité confidence pour évaluer la proximité thématique entre les termes sélectionnés. Cette mesure est le maximum entre deux probabilités conditionnelles. […] Nous avons appliqué l'algorithme de Louvain [35] pour identifier des groupes de termes délimitant des sujets. Enfin, nous avons généré la carte thématique pour chacun de ces deux corpus.

Dans cette méthode (tout comme la LDA, non utilisée ici car peu adaptée aux formats courts comme les tweets) « les sujets sont définis comme des "sacs de mots", déduits des statistiques d'apparition au sein des documents de termes issus d'une liste de mots clés définie à l'avance », et même si les mots sont ensuite liés aux thèmes, l'analyse reste tributaire d'une vision atomisée du lexique, qui ne prend pas en compte la phraséologie, la syntaxe, voire la dimension argumentative des termes. En outre, cette approche se distingue des méthodes textométriques dans lesquelles « le texte est caractérisé par ses mots par rapport à leur usage dans le corpus, le mot est caractérisé par ses cooccurrents, etc. » (Pincemin 2012). Cela n'est pas l'objectif de ce projet, qui est centrée sur l'analyse des communautés politiques, et la circulation des messages, et il est néanmoins très intéressant pour décrire l'influence des discours numériques sur les dynamiques politiques. Ce projet est donc

remarquable par son analyse topologique de la twittosphère, et sa mesure des traces numériques et des dynamiques des communautés. Il connaît néanmoins certaines limites si on le considère avec un prisme discursif, puisque le contenu des messages est analysé de manière assez systématique et peu contextuelle.

Un autre projet, porté par une équipe de linguistes-informaticiens, et d'historiens, s'inscrit dans l'objectif de décrire la campagne lors de son déroulement, avec une attention plus spécifique à la matérialité des corpus.

2.2. Mesure du discours et Hyperdeep : analyse de corpus et du vocabulaire

Comme indiqué sur le compte X du projet, http://mesure-du-discours.unice.fr est une plateforme web d'analyses statistiques de corpus politiques issue des travaux de recherches de l'UMR 7320 – BCL. Le projet combine des méthodes logométriques et le *deep learning*. Il croise donc fortement les intérêts et objectifs des projets menés à CY Cergy Paris université que nous décrirons dans un second temps.

Le site ne documente pas beaucoup son objectif général, mais le site de la bibliothèque universitaire de l'UCO (https://bu.uco.fr/ressource/mesure-du-discours) explicite les enjeux :

> Ce site web est co-développé par l'équipe "Logométrie" du laboratoire "Bases, Corpus, Langage" de l'Université de Nice et l'équipe SPARKS du laboratoire I3S localisé dans la même ville. Plus qu'un site, c'est un outil d'analyse des discours politiques français fonctionnant grâce à l'intelligence artificielle et la statistique textuelle.
> […]
> On y trouve les discours produits par les présidents successifs de la 5e République, ceux produits lors des dernières élections présidentielles (2007, 2012 et 2017) et plusieurs discours révolutionnaires (1789-1799). Chaque discours est accessible en texte intégral auquel s'ajoutent différents résultats d'analyse textuelle : les mots les plus spécifiques employés, les passages clés, les distributions statistiques de mots, les cooccurrences ou encore l'analyse par "deep learning" de l'orientation politique des discours.

L'essentiel de la base méthodologique et théorique de ce site est présenté dans l'ouvrage *L'Intelligence artificielle des textes. Des algorithmes à l'interprétation* (2021) publié sous la direction de Damon Mayaffre et Laurent Vanni. Plusieurs chapitres détaillent les aspects principaux et novateurs de l'approche. Dans leur article intitulé « Deep learning et description des textes. Architecture méthodologique », Laurent Vanni et Frédéric Precioso expliquent que « le deep learning montre aujourd'hui de nouveaux atouts et bénéfices pour l'Analyse de Données Textuelles (ADT) »

(p. 15), notamment parce que la boite noire que constitue souvent l'IA « s'ouvre peu à peu et libère une connaissance inédite des textes », complétant ainsi les méthodes traditionnelles et « offrant un champ de vision plus large et un parcours interprétatif plus fin ». La réflexion autour de l'IA concerne à la fois la nature des observables, et le statut accordé au texte et plus largement au langage (p. 17) :

> Avec les réseaux de neurones, il est possible d'observer des zones descriptives du texte plus ou moins longues, possiblement discontinues, définies par l'apprentissage du modèle. En deep learning il y a peu d'a priori sur le texte et ses observables : les mots, les lemmes, les codes grammaticaux sont encodés en valeurs numériques et le modèle recherche dans cette masse d'informations tout élément susceptible de lui permettre de converger vers une solution optimale.

L'encodage des mots dans les réseaux de neurones passe par le recours aux *embeddings* qui « à chaque mot un vecteur (potentiellement de grande dimension) de telle manière que la distance euclidienne des vecteurs reflète, sinon directement le sens des mots, toutefois leurs rapports et organisation sémantiques » (p. 22) : ceci offre en quelques sortes un changement de paradigme, puisque

> l'approche fréquentielle statistique classique donne aux mots une valeur commune pour toutes ses occurrences, quels que soient ses contextes. […] Avec la convolution, l'approche passe du fréquentiel au séquentiel, chaque mot est pris dans son contexte et engendre une représentation vectorielle particulière liée au contexte. Même si le fréquentiel joue un rôle important dans l'apprentissage, cette fenêtre contextuelle coulissante garantit une analyse séquentielle et contextualisante des mots. (Vanni, Precioso 2021, p. 29)

Plus généralement, pour la problématique de l'article et de ce numéro, on notera un changement radical de la manière de concevoir les mots, qui sont « convertis » en vecteurs, et deviennent donc des données numériques, qui peuvent ensuite être traitées par des algorithmes.

Plus précisément, les auteurs expliquent (p. 33) que « pour obtenir une représentation pertinente des mots, une des solutions les mieux reconnues par la communauté est d'utiliser une méthode appelée Word2Vec, sorte de préapprentissage qui utilise différents modèles possibles comme CBOW ou Skip-Gram de (Mikolov *et al.* 2013). Cette méthode efficace prend en compte l'axe syntagmatique et l'axe paradigmatique [et] donne un moyen efficace de fixer la représentation des mots avec un état compréhensible pour l'homme basé sur la cooccurrence des mots », avec finalement (p. 34) une « représentation des mots par leurs profils cooccurrentiels ». Pour pouvoir aboutir sur une forme interprétable pour l'humain, les auteurs introduisent une notion supplémentaire à leur modèle, la *déconvolution* : cette méthode

« agit comme un décodeur pour les réseaux à convolution. Adaptée au texte cette déconvolution propose une nouvelle manière de lire les textes sous une forme abstraite proposée par le deep learning » (p. 41). Les auteurs appellent ce processus « text deconvolution saliency (TDS) » qu'ils définissent comme « un ensemble de processus algorithmiques qui visent à inverser les effets de la convolution » (p. 41) : ils expliquent également que « linguistiquement cela se traduit par une mise en valeur de certains mots, expressions ou patterns que le réseau utilise pour prendre sa décision (classification) »[2] (p. 46).

Le chapitre suivant, intitulé « Littérature et intelligence artificielle » et écrit par Étienne Brunet, Ludovic Lebart, Laurent Vanni, évoque des enjeux plus généraux du recours en IA dans l'analyse des textes, et détaillent surtout le type de rapport qui est donné à l'analyste avec les résultats (pp. 74-75) :

> lorsque les techniques de l'Intelligence artificielle sont mises en œuvre, la reconnaissance (ou identification) n'est pas absolue. Le processus de classification aboutit à un continuum de valeurs échelonnées de 0 à 100. Cette approximation en pourcentages est assez habituelle dans les conclusions statistiques, et particulièrement dans les méthodes éprouvées de la lexicométrie. Mais il faut ici souligner une différence essentielle dans l'approche des faits. L'habitude s'est imposée dans les études lexicométriques de constituer un corpus en rassemblant des textes que l'on oppose les uns aux autres en s'aidant de la « norme » interne représentée par l'ensemble du corpus. Il y a certes des avantages pratiques à procéder de la sorte mais cela ne va pas sans bousculer la vraie démarche statistique qui exclut radicalement l'échantillon de la population. Pour savoir si un échantillon (ou un texte) peut ou non appartenir à la population (ou corpus), les tests autorisés n'admettent pas sa présence, même proportionnellement faible, dans les données de référence, dites aussi données d'apprentissage. La procédure que suit en particulier le deep learning respecte intégralement cette distinction, et dénonce immédiatement les forfaitures, en parlant de façon assez dérogatoire d'apprentissage par cœur (overfitting). Quand le texte à examiner figure par mégarde ou ignorance dans les données d'apprentissage, un seuil scandaleux de reconnaissance signale le non-respect du principe de séparation. Dans la même situation, la lexicométrie traditionnelle enregistre peu de différence dans les résultats, que le texte à examiner soit extérieur ou non au corpus. On peut en conclure a priori que l'approche du deep learning, étant plus exigeante et plus sévère dans les conditions de traitement, laisse espérer une sensibilité plus grande et plus exacte à la spécificité des textes.

[2] Ou encore : « Le TDS est donc un nouvel outil qui attribue un poids à chaque mot, basé sur un calcul non linéaire (lié aux fonctions d'activation de chaque neurone, elles-mêmes non linéaires) et que la statistique classique ne peut engendrer. Avec la convolution, une même forme graphique prend un poids différent suivant son contexte, là où la statistique ne propose qu'un seul score de spécificité (ou z-score) pour chacune des classes ou qu'un profil cooccurrentiel général unique dans le corpus ».

On reconnaît donc une complémentarité des approches, mais également une différence de fond en ce qui concerne la reconnaissance de manière « non absolue » en pourcentage (pour le *deep learning*) d'un côté, ou la proximité/distance des textes en fonction de la norme établie par l'ensemble des textes (le corpus, dans la lexicométrie/statistique textuelle).

Ces technologies sont ensuite mobilisées en ligne, comme dans la Figure 1 avec l'analyse du mot « combat » dans l'élection présidentielle 2017 :

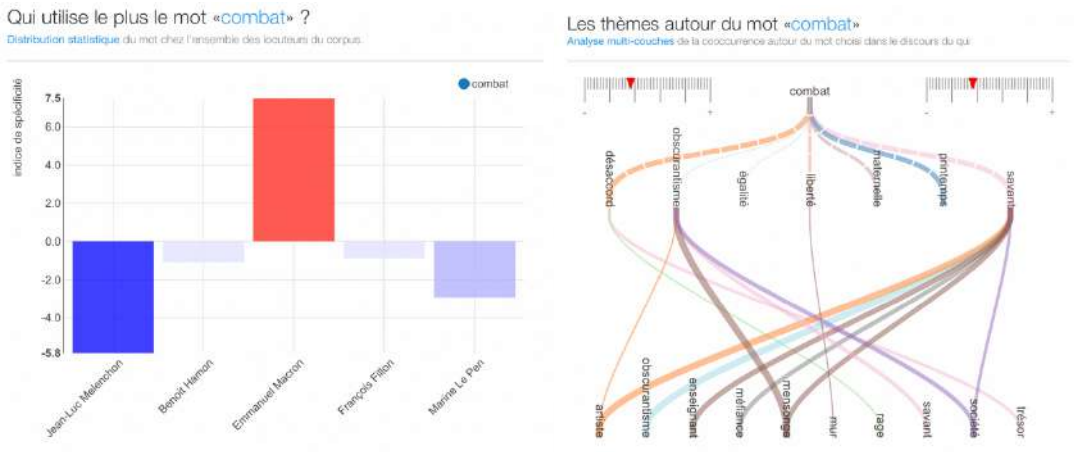

Figure 1
Analyse du mot « combat » dans Mesure du Discours (présidentielle 2017).

Les préoccupations de ce site rejoignent un certain nombre des enjeux de projets que nous avons également menés, centrés en particulier sur les liens entre lexique, idéologies, et argumentation politique.

3. #Idéo2017 et PolitiQuiz, deux projets ouverts et interactifs pour explorer les mots du politique dans la transformation numérique

À partir du second semestre 2016, et dans le cadre d'un projet financé par CY Fondation, nous avons mis en place un projet de recherche dont l'objectif était de rendre les outils et techniques de l'analyse du discours numérique au plus grand nombre.

3.1. Interactivité et acquisition de nouvelles connaissances

Le projet #Idéo2017, centré sur les élections présidentielles 2017 a permis la mise en ligne d'une plate-forme d'analyse, en temps réel, des tweets des candidats, avec plusieurs types de métriques et visualisations de leurs corpus. La chaîne de traitement proposée était celle présentée en Figure 2 :

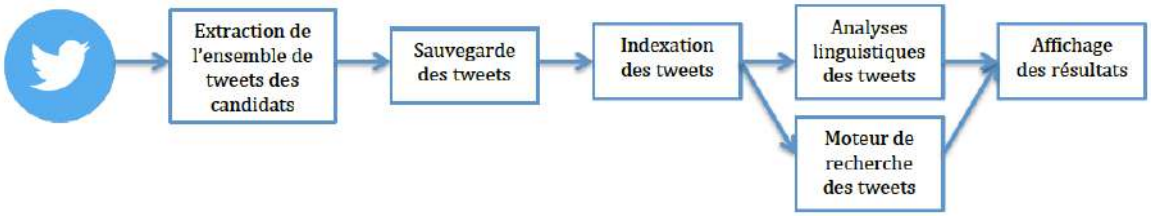

Figure 2
Chaîne de traitement de la plateforme #Idéo2017.

En ligne, l'interface visible par l'utilisateur se composait de trois parties, permettant de lancer des analyses linguistiques de mots, de corpus des candidats, ou d'accéder à un moteur de recherche (Figure 3) :

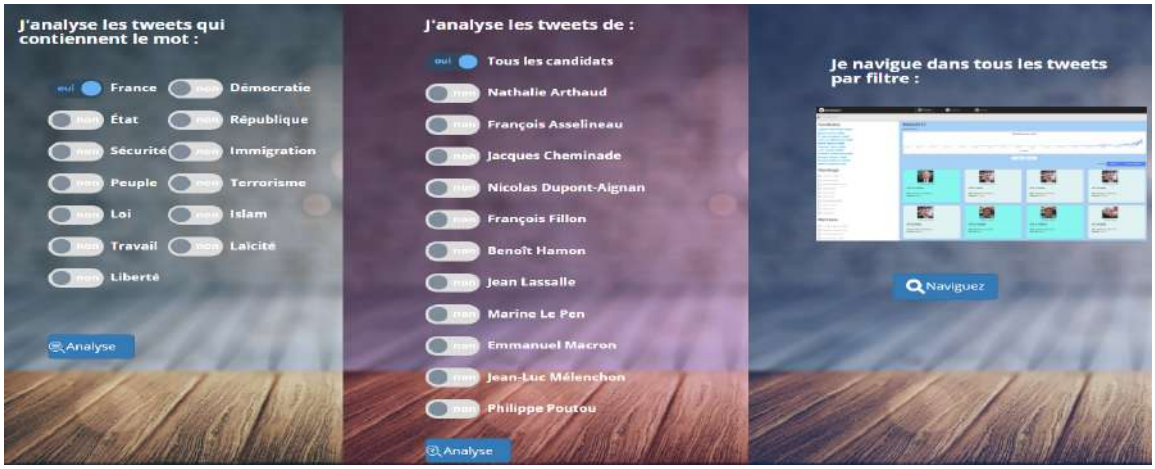

Figure 3
Plateforme #Idéo2017 côté utilisateur.

Plusieurs publications ont déjà résumé les enjeux et concepts du projet #Idéo2017, notamment dans Longhi (2019) :

> L'analyse "J'analyse les tweets qui contiennent le mot..." permet à l'utilisateur de choisir un mot parmi les 13 mots qui sont souvent employés dans les débats politiques (ALDUY 2017). Cette entrée donne accès à quatre analyses possibles : l'usage de ce mot par les différents candidats (sur/sous-emploi et fréquences de la forme exacte), les mots associés à ce mot (analyse de similitudes, basée sur les cooccurrences entre les mots), l'emploi de ce mot et ses dérivés par les différents candidats […] et le nuage de mots. Ces analyses sont en fait des résultats produits grâce au code du logiciel d'analyse textuelle Iramuteq, issus de calculs qui portent dans le logiciel des noms plus techniques (l'analyse de similitude devient par exemple les mots associés à ce mot).
> L'analyse "J'analyse les tweets de [candidat]" permet à l'utilisateur de choisir un candidat parmi les 11 candidats (ou le corpus global des 11 candidats) afin d'analyser ses tweets via les techniques suivantes : les mots les plus utilisés, les thématiques (issues de la méthode Reneirt), les relations entre les mots, le

> nuage de mots, les spécificités des différents candidats (possible si l'utilisateur a choisi d'analyser tous les candidats en même temps).
> Enfin, le moteur de recherche permet à l'utilisateur de faire des recherches sur toute la base des tweets, grâce à un outil appelé ElastricSearch.

Ces analyses, comme dans Mesure du discours, permettent à l'utilisateur d'accéder à des résultats d'analyses textométriques, et de les rendre visibles dans le cadre d'une interaction avec l'interface. Il est ici davantage encore question de centrer la plateforme sur des « objets discursifs » (Longhi 2008, 2015) dans le but de pouvoir comparer la dimension idéologique des discours des candidats, et le travail réalisé sur la matérialiste des discours qu'ils tiennent.

Dans ce cadre du projet PolitiQuiz[3], nous avons souhaité contribuer à la compréhension des élections régionales 2021 en Île-de-France, puis aux élections présidentielles 2022 en France. Ce projet propose des questions et analyses permettant de mieux connaître les discours des candidates, leur style, leurs thèmes, et leurs programmes. Mais au-delà du simple quiz, PolitiQuiz offre aux utilisateurs l'accès aux technologies développées dans le cadre de l'institut des humanités numériques (IDHN) de CY.

3.2. Des questions pour mieux comprendre les discours des candidats

L'application PolitiQuiz proposait initialement trois grandes thématiques : les connaissances politiques, l'attribution d'auteur, et les grandes thématiques de campagnes (sécurité, santé, etc.). Dans la version de 2022, le choix a été simplifié à deux catégories, davantage centrées sur la thématique politique que sur le type de traitement. On distinguait ainsi la « culture générale » d'une partie concernant « Les programmes en question » (voire Figure 4) :

[3] Plusieurs participant-e-s ont collaboré au projet PolitiQuiz, dans ses différentes phases : Jérémy Demange, Ingénieur d'études (IGE), a été le maillon central du projet sur le plan technique et multimédia. Des stagiaires (de DUT ou de Master) ont été impliqués dans la première phase (design, questions) ; des doctorantes (Lise Pernet et Rose Moreau Raguenes) ont également participé à la 2nde phase (élection 2022) en travaillant notamment sur les programmes et sur la dimension participative. Cette équipe a donc participé à ce qui est résumé dans les sections 3.2 et 3.3.

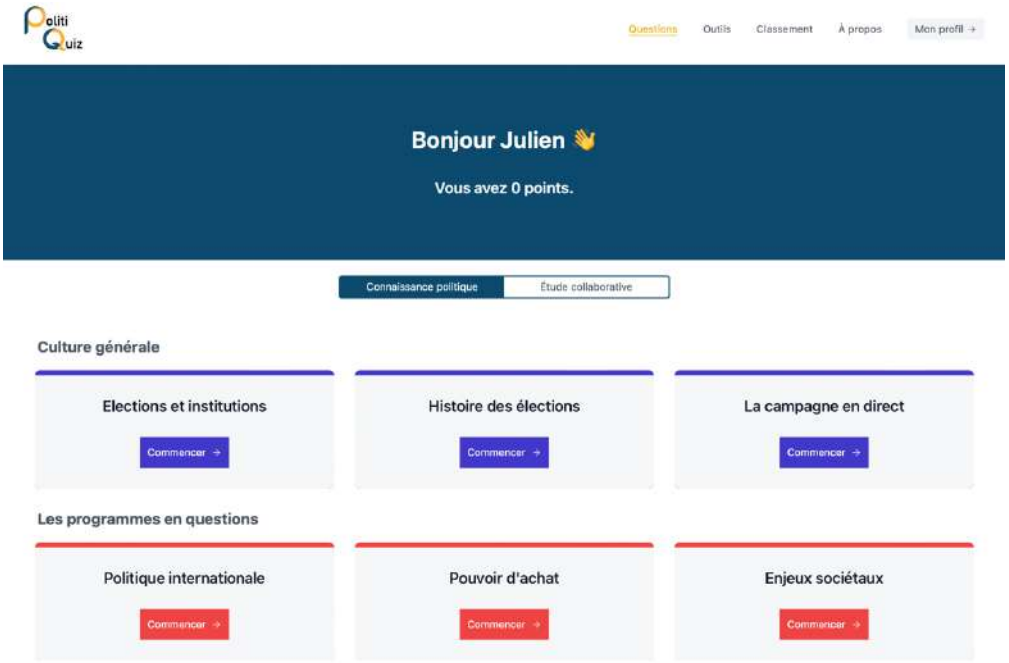

Figure 4
Plateforme PolitiQuiz côté utilisateur.

Les questions ont été produites à partir de l'analyse des tweets des principaux candidats à l'élection (régionale puis présidentielle). Pour cela, nous avons utilisé différents outils ou logiciels (comme Iramuteq ou Tropes) et méthodes d'analyses afin de pouvoir créer de nouvelles questions. Il s'agissait en effet d'extraire des caractéristiques thématiques, sémantiques, du corpus de référence initialement constitué, pour produire des questions adaptées à l'élection en cours. Par exemple, en utilisant le « calcul de spécificité » du logiciel Iramuteq, nous avons mis en valeur les mots sur- et sous-employés par tel ou tel candidat. À partir de ces résultats, des questions ont été rédigées (par l'équipe du projet et une stagiaire recrutée pour chacune des élections), pour permettre aux utilisateurs de tester leurs connaissances sur les mots les plus utilisés par les différents candidats, ou de comparer les discours des candidats, et donc leurs propositions, comme la Figure 5 s'illustre :

Figure 5
Exemple de question posée sur PolitiQuiz.

Le travail a été réalisée de manière collaborative comme indiqué en Figure 6 :

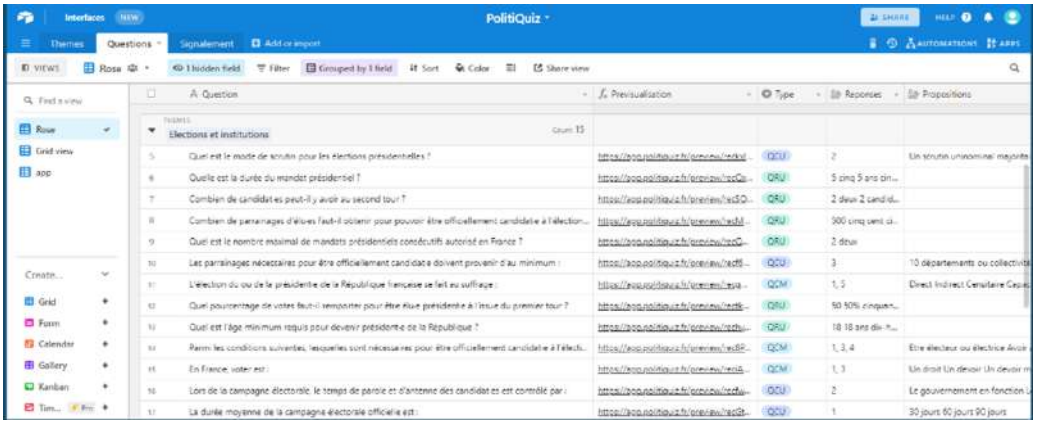

Figure 6
Document collaboratif de saisie des questions
(complété par les stagiaires et participant-e-s au projet).

Notons que l'application possédait un algorithme auto-adaptatif, qui adaptait les questions en fonction des réponses et proposait ensuite des questions plus ou moins complexes. En fonction des questions, nous avions attribué à chacune d'elle un niveau de difficulté sur une échelle de un à trois. En fonction de la progression de l'utilisateur, l'outil proposait une pause d'un ou deux jours pour faire travailler la mémoire, puis de nouvelles questions inédites apparaissaient à nouveau sur l'application.

Cette dimension pédagogique vient du fait que le projet PolitiQuiz avait des liens avec un autre projet Open Source d'une start-up d'État : Pix.fr. Le code de Pix a été redéveloppé (par Jérémy Demange) en ajoutant des questions, un classement et une interface aux couleurs du PolitiQuiz. L'application avec ces modifications permettait d'offrir une interface intuitive et facile à prendre en main, qui embarque un certain nombre de technologies (Figure 7) :

Figure 7
Page de création de compte dans PolitiQuiz.

Le site s'appuyait sur différentes innovations scientifiques, en les rendant accessibles, et ludiques, par exemple :

- pour les questions de la catégorie « attribution d'auteur », l'algorithme d'attribution d'auteur développé dans le cadre du projet CHEMI « IRITA », a été combiné à un processus de génération de texte par l'usage du modèle GPT-Neo ;
- pour les questions liées aux thématiques politiques, nous avons préalablement eu recours aux classifications statistiques des corpus. En effet, un traitement d'un large corpus de tweets des candidats a été établi avec le logiciel de statistique textuelle Iramuteq, afin de caractériser les mondes lexicaux (classes) qui rendent compte des grandes thématiques : éducation, sécurité, etc.

Une page de classement est également accessible, permettant de comparer ses résultats avec ceux des autres utilisateurs de l'application. En utilisant le principe du jeu, du classement, voire du concours, nous espérions sensibiliser les électeurs aux différents enjeux du scrutin, tout en leur offrant des connaissances et informations, de manière interactive et ludique.

3.3. Science participative et linguistique populaire : l'analyse des nominations

En plus de la dimension ludique et analytique, nous avons, à la fin de l'élection 2022, pris l'initiative d'introduire une nouvelle composante, intitulée « Étude collaborative » (figure 8) :

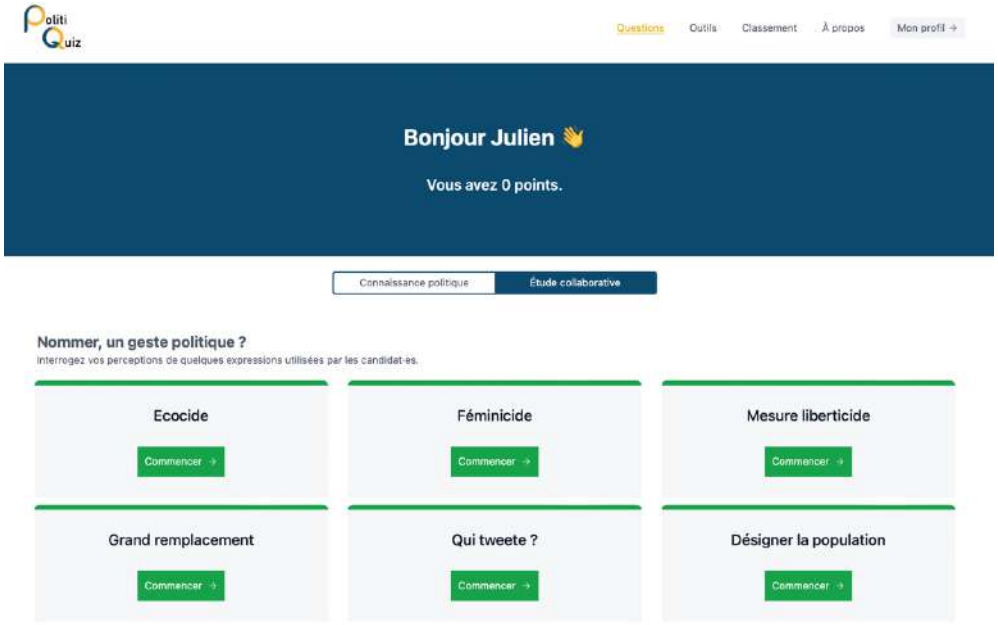

Figure 8
Étude collaborative au sujet de la nomination (« Nommer, un geste politique » ?).

Il s'agissait dans cette rubrique d'« interroger [les] perceptions de quelques expressions utilisées par les candidat.e.s ». L'intitulé, « Nommer, un geste politique », inscrit ce questionnement dans le domaine de la nomination et de l'acte de nommer, qui constitue une voie d'entrée privilégiée dans l'analyse sémantiques des discours. Des termes comme « écocide », « féminicide », ou des expressions comme « Grand remplacement », étaient questionnées.

4. Articuler « recherche-action » et réflexivité : réinterroge le concept de style au prisme des formations discursives et des genres de discours

La réalisation de tels projet, en prise direct sur la campagne, nécessite la production de résultats rapides, et la mise en évidence de tendances sans forcément pouvoir les approfondir sur le moment. Ces corpus, résultats préliminaires, et analyses, sont autant d'hypothèses qui peuvent ensuite être approfondies par le chercheur. En particulier, notre entrée lexicale dans l'analyse du discours politique permet de comparer le vocabulaire des candidats, mais aussi leur style.

4.1. Comparaison des spécificités : le vocabulaire comme entrée dans le style en politique

Lors de ces deux élections, certain-e-s même candidat-e-s ont été présents, et cela nous a invité à questionner la permanence du style de ces personnalités, ainsi que la cohérence de leur discours à cinq années d'intervalle. La question du contexte est également à considérer, puisque les enjeux peuvent évoluer d'une élection à l'autre. Nous avons donc choisi de comparer trois personnalités, lors de ces deux élections : Emmanuel Macron, Marine Le Pen et Jean-Luc Mélenchon. Emmanuel Macron étant président en 2022, sa communication diffère de celle d'une communication de campagne classique, ce qui nous a conduit à adapter la collecte comme indiqué sur la figure 9 :

	auteur	size
0	EM1	2612
1	EM2	298
2	JLM1	3722
3	JLM2	1000
4	MLP1	2440
5	MLP2	1000

Figure 9
Composition du corpus.

Nous n'avons que 298 tweets d'Emmanuel Macron sur cette période, et nous avons ensuite limité à 1000 le nombre de tweets de Jean-Luc Mélenchon et Marine Le Pen. Le recours au logiciel Iramuteq est utile ici, puisqu'il permet d'encoder des variables, comme l'auteur ou la date, et de calculer ce qui est spécifique de telle ou telle partie de corpus (figure 10) :

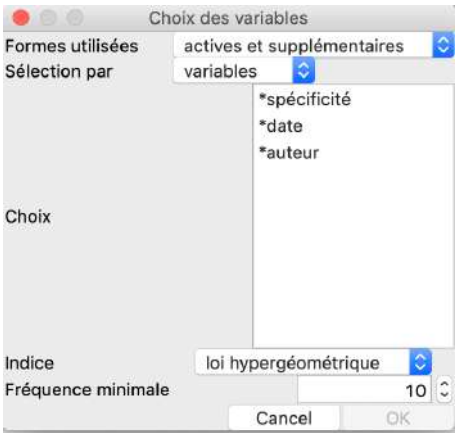

Figure 10
Spécificités analysables dans Iramuteq.

Ceci est rendu possible par l'encodage du corpus (Figure 11), où on peut d'ailleurs distinguer l'auteur (Emmanuel Macron par exemple) du candidat spécifique (appelé ci-dessous spécificité, « EM2 » soit Emmanuel Macron de 2022) :

```
**** *auteur_EM *date_24_03_2022 *spécificité_EM2
Pour éviter une crise alimentaire, il nous faut réagir. J'ai souhaité lancer, en lien
direct avec l'Union africaine, l'initiative FARM : Food and Agriculture Resilience
Mission.

**** *auteur_EM *date_24_03_2022 *spécificité_EM2
La résolution humanitaire portée par la France et le Mexique a été votée à une très large
majorité à l'Assemblée générale des Nations unies : 140 voix pour, 5 voix contre. Cela
montre l'isolement que nous maintenons sur la Russie.

**** *auteur_EM *date_24_03_2022 *spécificité_EM2
Avec nos Alliés, nous allons continuer à fournir des armements, dans un cadre d'efficacit
complète, avec une ligne rouge : ne pas être co-belligérants.
```

Figure 11
Balisage du corpus.

Sur la base de ce découpage, le chercheur peut comparer les spécificité (sur- et sous-emploi) en choisissant des termes qui l'intéressent. Nous avons choisi d'analyser des mots en lien avec l'éducation et avec l'écologie.

Concernant l'éducation, nous avons sélectionné les lemmes « élève », « éducation », « école », « enseignement » et « enseignant » et procédé à une représentation graphique de leur sur- ou sous-emploi dans les parties de corpus (Figure 12) :

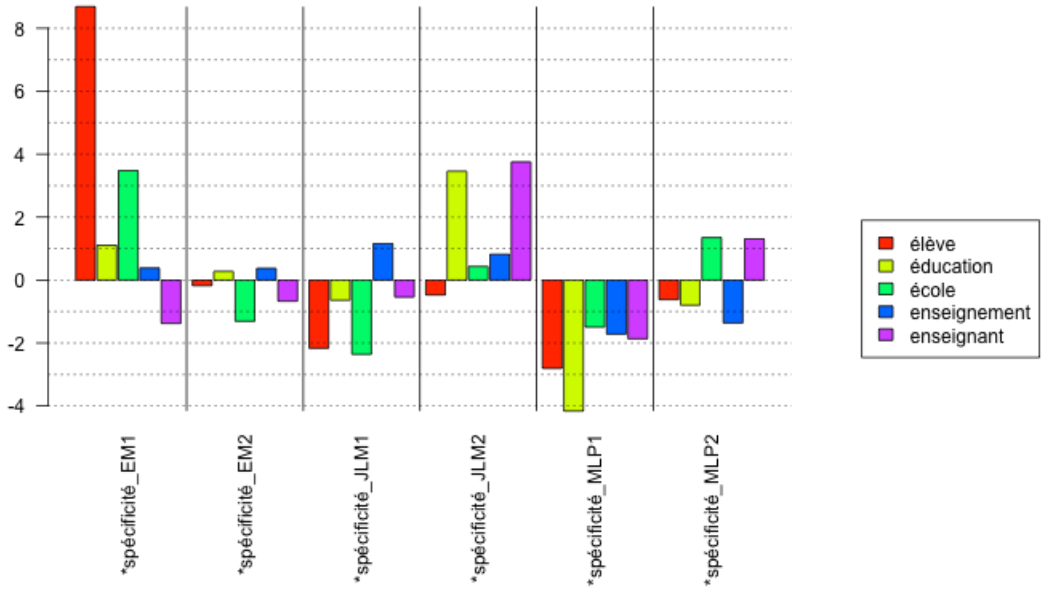

Figure 12
Spécificités de termes liés à l'éducation.

On note que les candidats qui sur-utilisent ces termes sont Emmanuel Macron 2017 (« élève » et « école » notamment ainsi que « éducation », Jean-Luc

Mélenchon 2022 (avec « éducation » et « enseignant » en particulier), et Marine Le Pen 2022 dans une moindre mesure (avec « école » et « enseignant »). Il faut bien sûr ensuite retourner au corpus pour mener l'analyse complète (qui n'est pas le propos de cet article) mais on voit bien comment l'accès à ce corpus réalisé en « temps réel » lors de campagnes peut ensuite donner lieu à des analyses sociodiscursives plus complexes. La même chose peut être réalisé autour des mots de l'écologie (Figure 13), que l'on peut utiliser pour montrer la prégnance de « écologique » chez Jean-Luc Mélenchon (et son sous-emploi chez Emmanuel Macron 2017 et Marine Le Pen), l'emploi de « environnemental » et « environnement » chez Emmanuel Macron en 2017, remplacé par « climatique » en 2022 :

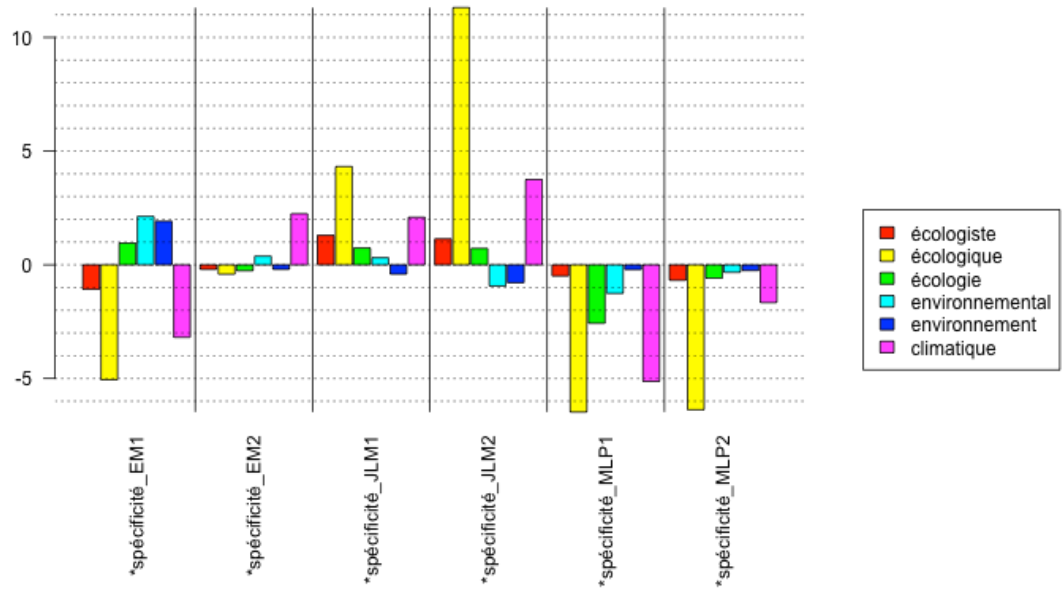

Figure 13
Spécificités de termes liés à l'écologie.

L'analyse des spécificités a donc l'avantage de pouvoir « mesurer » les corpus non pas à travers le calcul des fréquences, mais en lien avec l'importance du vocabulaire de chaque partie relativement à l'ensemble du corpus.

Cette analyse comparative peut être étendue, en termes de variables et de perspectives, grâce au recours au *deep learning*.

4.2. Le deep learning *et l'adaptation à plusieurs niveaux d'analyse*

Comme nous l'avons expliqué à propos de « Mesure du discours », le recours au *deep learning* permet un nouveau type de traitement, qui permet notamment de classifier les résultats en fonction de certaines problématiques. Dans un travail mené à propos de l'attribution d'auteur (Lam, Demange,

Longhi 2021), le paramétrage suivant (Figure 14) a été établi, pour donner les résultats optimums sur l'analyse d'un corpus de tweets :

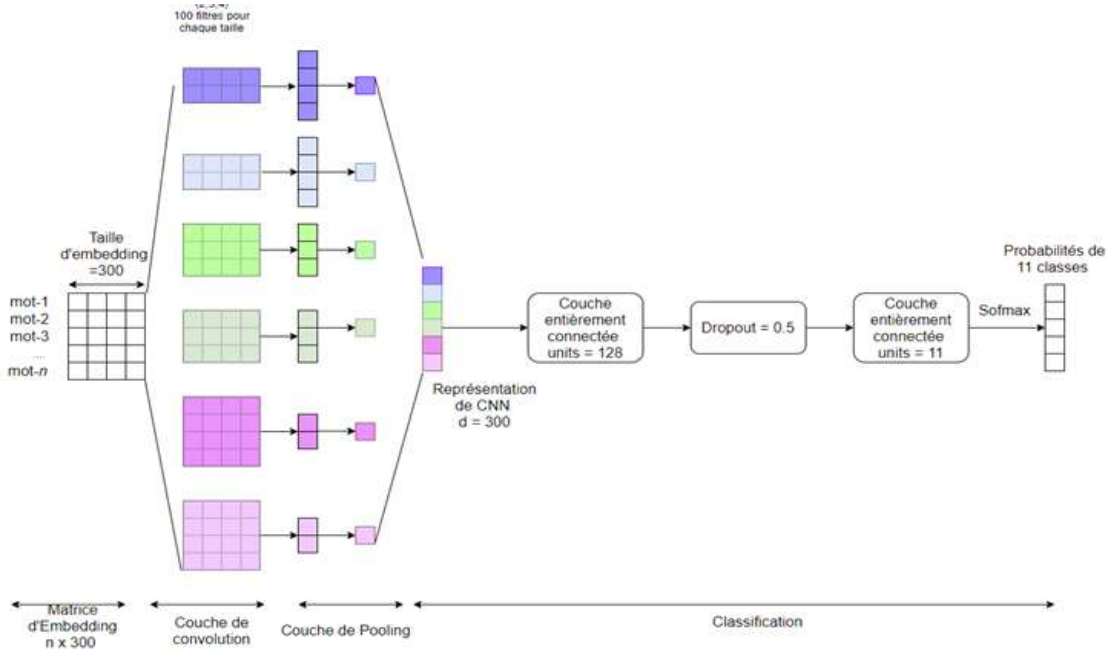

Figure 14
Choix du modèle et paramétrage du classifieur.

Ce modèle a été appliqué au corpus précédemment étudié avec Iramuteq, afin d'observer si le classifieur prédit correctement l'auteur du message. Nous représentons les résultats dans la matrice de confusion en Figure 15, qui confronte les classes prédites (en ordonnées) aux vraies classes (en abscisse) :

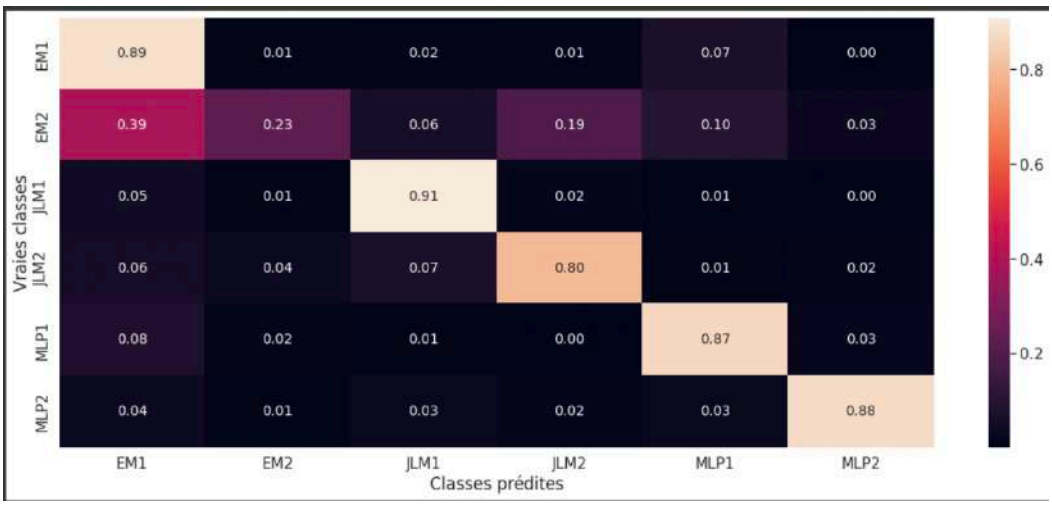

Figure 15
Matrice de confusion (classes prédites/vraies classes) avec les 6 variables.

On observe ainsi que la plupart des résultats sont corrects voire bons (diagonale de la figure), avec par exemple un score de 0,91 de bonne

prédiction (précision) pour Jean-Luc Mélenchon 2017, ou 0,89 pour Marine Le Pen 2022. Mais la ligne EM2 (Emmanuel Macron 2022) montre une disparité forte avec le reste des résultats, puisque le score est de seulement 0,23 pour les bonnes prédictions (alors que le classifieur prédit 0,39 pour EM1, et 0,19 avec Jean-Luc Mélenchon 2022). À ce premier niveau, les « confusions » du classifieur sont intéressantes, puisque cela nous permet d'anticiper sur d'éventuelles proximités ou similarités entre candidats.

Comme préconisé par Vanni et ses collègues, le principe de déconvolution est ici utile pour rendre ces classifications intelligibles. Dans les deux figures suivantes, on observe des erreurs de classification.

Dans l'exemple suivant (Figure 16), le classifieur indique EM1 (Emmanuel Macron 2017) alors que le vrai label est MLP1 (Marine Le Pen 2017) :

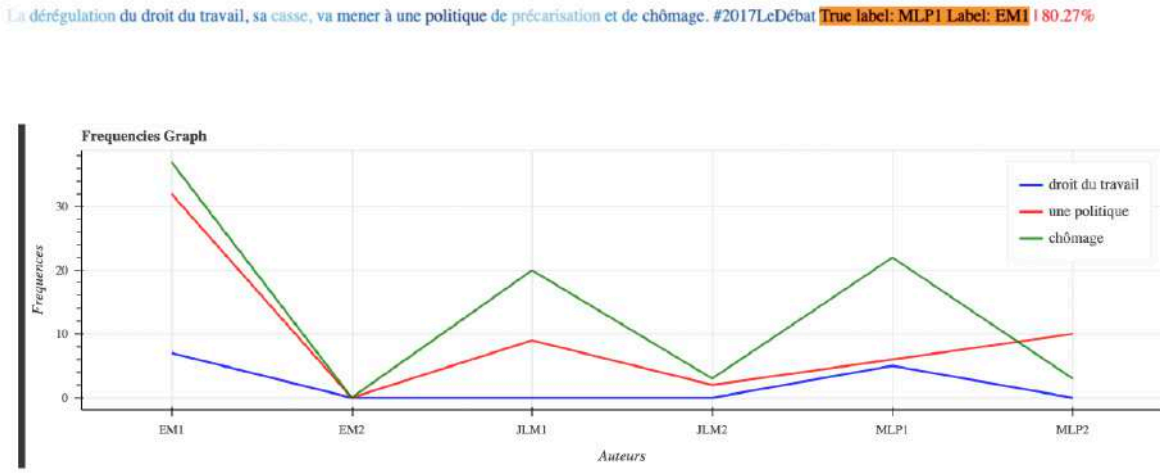

Figure 16
Déconvolution à propos d'un exemple de MLP1 attribué à EM1.

Les expressions/termes « droit du travail », « une politique » et « chômage » sont distingués comme caractéristiques du discours de d'Emmanuel Macron en 2017. Or dans ce message de Marine Le Pen, on a la coïncidence de ces trois marqueurs : cela explique la confusion du réseau, qui a attribué le message à EM1.

Dans le second exemple (Figure 17), le classifieur fait cette fois une confusion entre EM1 et JLM1 : à partir de la co-présence de « tribune », « un moment de », « débat » et « convaincre », les classifieur penche pour EM1 alors qu'il s'agit de JLM1 :

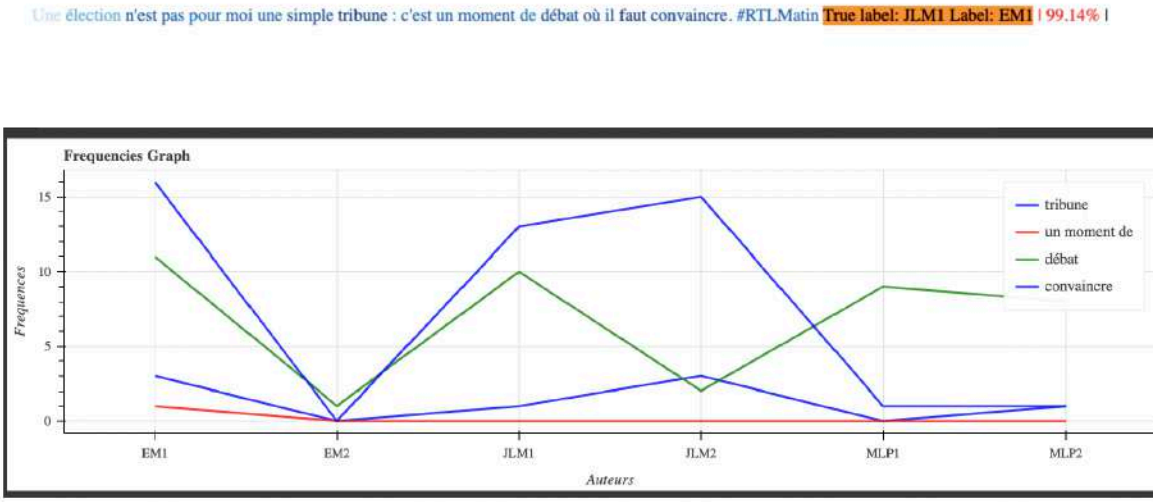

Figure 17
Déconvolution à propos d'un exemple de JLM1 attribué à EM1.

Ces confusions sont utiles pour l'analyste du discours intéressé par le discours politique, car elles révèlent des enjeux politico-discursifs importants : proximité entre candidats, adaptation des discours aux contextes, proximités lexicales sur certains thèmes, etc. Dans le cas qui nous intéresse, une particularité autour du candidat EM2 tient au fait que dans la campagne 2022, il est le président sortant, et a donc un comportement discursif très différent. Aussi, pour compléter l'analyse, nous avons cherché les comptes qui pourraient se faire le relai du « candidat Macron », puisque le compte officiel coïncide avec celui du président (ce qui entretiendrait une certaine confusion des genres).

4.3. Prise en compte du contexte politique et élargissement du corpus

Nous avons ainsi repéré deux comptes susceptibles de « faire campagne » en lieu et place de celui du président : « En Marche » (EMA), et « avec vous », caractérisé par « Emmanuel Macron avec vous ! » (Figure 18) :

Figure 18
Description des 2 comptes ajoutés pour la campagne 2022.

Pour savoir si cela a une influence sur les résultats, et l'analyse par le modèle, on peut lancer une nouvelle analyse et observer dans la Figure 19 les prédictions du modèle :

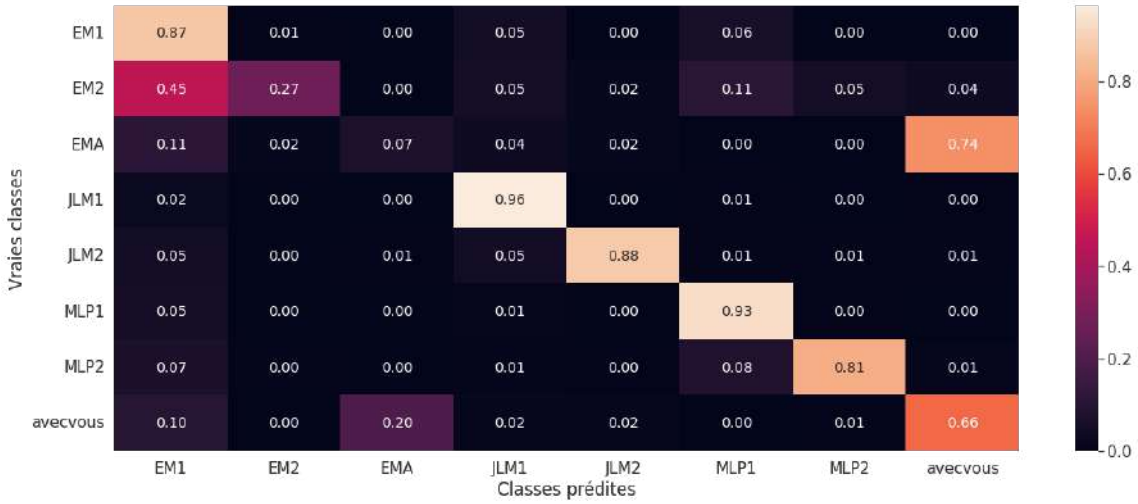

Figure 19
Matrice de confusion (classes prédites/vraies classes) avec les 8 variables.

La ligne EMA est particulièrement intéressante, puisque les prédictions sont extrêmement mauvaises (0,07), et les erreurs les plus importantes sont avec EM1 (0,11) et surtout avec avecvous (0,74). Les score pour avecvous sont très moyens (0,66), et les erreurs sont essentiellement avec EMA (0,20) et EM1 (0,10). Les erreurs ont donc un sens, puisqu'il y a des confusions très fortes, dans les différentes configurations, entre EMA, avecvous, et EM1. Le style d'argumentation électoral macronien se transpose donc ailleurs en 2022 pour ne pas se confondre avec la parole présidentielle, et des relais sont mis

en place pour conserver une continuité avec la posture argumentative du contexte de la campagne.

5. Conclusion

L'évolution des technologies numériques a profondément transformé la manière dont nous échangeons des idées politiques. La montée en puissance des réseaux sociaux, des blogs et des forums de discussion en ligne a offert une plate-forme de plus en plus accessible pour les organisations politiques de communiquer et d'interagir avec les électeurs. L'une des principales évolutions du discours politique dans le contexte numérique est l'accent mis sur l'interaction et la participation. Les candidats et partis politiques utilisent les réseaux sociaux pour atteindre un public plus large et pour mobiliser leurs sympathisants lors des élections. Cette évolution du discours politique n'est pas sans ses défis : les plateformes de médias sociaux ont souvent été critiquées pour favoriser la polarisation et l'incitation à la haine. Aussi, l'analyse du discours politique numérique est devenue de plus en plus importante en raison de l'importance croissante d'Internet et des médias sociaux dans le débat public et la formation de l'opinion. Les nouveaux défis de l'analyse du discours politique numérique concernent notamment la quantité de données (qui sont souvent bruyantes et contiennent des informations redondantes ou peu pertinentes), la polarisation, la désinformation, et la difficile distinction entre les faits, les opinions et les rumeurs, notamment dans des messages courts. L'analyse du discours politique numérique est un domaine en constante évolution, avec de nouveaux défis à relever : les chercheurs doivent être capables d'adapter leurs méthodes et leurs outils pour répondre à ces défis et produire des analyses précises et pertinentes.

Dans cet article, nous avons présenté différents projets qui, au cours des élections présidentielles françaises 2017 et 2022, ont contribué à rendre accessibles les méthodes, outils et résultats de la recherche auprès d'un plus grand nombre. En détaillant notamment certains enjeux scientifiques autour de la classification des candidats, nous avons pu montrer l'intérêt scientifique et citoyen de tels projets, qui renouvellent tout autant la manière de faire de l'analyse du discours, que l'analyse des observables eux-mêmes, et les critères mis en œuvre pour faire émerger des résultats. La transformation numérique est donc un moyen de renouveler les terrains, méthodes, et modes de diffusion, de l'analyse du discours, et devient une opportunité pour observer les évolutions et mutations de la communication politique.

Bionote: Julien Longhi is *Professeur des universités* at CY Cergy Paris Université. He holds a PhD in discourse analysis at the University of Clermont-Ferrand II, and a *Habilitation à diriger des recherches* at CY Cergy Paris Université. His research focuses on the question of ideologies in political discourse, and on the use of discourse analysis to explore corpora. He is the director of the Digital Humanities Institute at CY, and has published several books and articles in the fields of digital discourse analysis, digital humanities, and tooled linguistics.

Author's address: julien.longhi@cyu.fr

Acknowledgements: We would like to thank CY Fondation and the National Research Agency (ANR) for their precious financial support to the current research.

Références bibliographiques

Brunet E., Lebart L. et Vanni L. 2021, *Littérature et intelligence artificielle*, in Mayaffre D. et Vanni L. (éds.), *L'intelligence artificielle des textes*, Honoré Champion, Paris, pp. 73-130, coll. Lettres Numériques.

Gaumont G., Panahi M. et Chavalarias D. 2018, *Reconstruction of the socio-semantic dynamics of political activist Twitter networks -Method and application to the 2017 French presidential election*, in "PLOS ONE", https://journals.plos.org/plosone/article ?id=10.1371/journal.pone.0201879.

Lam T., Demange J. et Longhi J. 2021, *Attribution d'auteur par utilisation des méthodes d'apprentissage profond*, in « EGC 2021 Atelier DL for NLP : Deep Learning pour le traitement automatique des langues ». https://hal.science/hal-03121305/document

Lebart L. et Salem A. 1994, *Statistique textuelle*, Dunod, Paris.

Longhi J. 2008, *Objets discursifs et doxa : essai de sémantique discursive*, L'Harmattan, Paris.

Longhi J. 2015, *L'acte de nommer comme constitution de formes : discursivité de la production du sens*, in « Langue française » 188 [4], pp. 121-136.

Longhi J. 2016, *Le tweet politique efficace comme mème textuel : du profilage à viralité*. in « Travaux de linguistique » 73 [2], pp. 107-126.

Longhi J. 2019, *Le projet #Idéo2017 : Quelles implications du/de la chercheur-e en tant qu'acteur-trice potentiel du changement social ? Exemplification à partir du discours politique numérique*, in « Cahiers de Linguistique » 44 [2], pp. 103-116.

Longhi J. 2020, *Proposals for a discourse analysis practice integrated into digital humanities: theoretical issues, practical applications, and methodological consequences*, in "Languages" 5 [1], 5. https://doi.org/10.3390/languages5010005

Mayaffre D. et Vanni L. (éds.) 2021, *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Honoré Champion, Paris.

Mikolov T., Sutskever I., Chen K., Corrado G.S. and Dean J. 2013, *Distributed Representations of Words and Phrases and Their Compositionality*, in Burges C., Bottou L., Welling M., Ghahramani Z. and Weinberger K. (eds), *Advances in Neural Information Processing Systems*, Vol. 26, Curran Associates, Inc., Red Hook pp. 3111–3119.

Pincemin B. 2012, *Hétérogénéité des corpus et textométrie*, in « Langages » 187, pp. 13-26. https://doi.org/10.3917/lang.187.0013.

Vanni, L. et Precioso, F. (2021). *Deep learning et description des textes. Architecture méthodologique*, in Mayaffre D. et Vanni L. (éds.), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Honoré Champion, Paris, pp. 73-130, coll. Lettres Numériques.