Que peuvent les algorithmes de plongement de mots pour l'analyse sociologique des textes ?

Analyser les discours et caractériser les locuteurs des plateformes « Grand Débat National » et « Vrai Débat »

Philippe SUIGNARD[1]

Caroline ESCOFFIER[2]

Lou CHARAUDEAU[3]

Mathieu BRUGIDOU[4]

TITLE

How can word embedding algorithms contribute to the sociological analysis of texts?
To analyze speeches and characterize speakers of the "Grand National Debate" and "Real Debate" platforms

RÉSUMÉ

Dans cet article, nous nous proposons de contribuer à l'évaluation de l'apport des algorithmes dits de « plongement de mots » à l'analyse sociologique des textes : d'une part, en confrontant les résultats des analyses sémantiques de ces algorithmes aux approches maintenant bien connues des analyses de données textuelles ou de textométrie ; d'autre part, en s'intéressant à ce qui constitue un des principaux obstacles à l'analyse sociologique du web : la difficulté à caractériser sociologiquement les auteurs des énoncés issus du web. Pour cela, nous analysons les énoncés issus de plateformes de « civic tech » – plateforme gouvernementale, le « Grand Débat National », et sa riposte politique et algorithmique proposée par un collectif de Gilets jaunes, le « Vrai Débat ». Un troisième corpus issu de la plateforme « Entendre la France », au design identique à celui du Grand Débat National et par ailleurs documenté en termes de propriétés socio-politiques, nous permettra de caractériser les locuteurs en fonction de leurs discours et de tenter de prédire par des approches de machine learning des « pseudos propriétés » affectées aux locuteurs du Grand Débat National.

Mots-clés **:** *plongement de mots, analyse des données textuelles, civic tech, débat public, Gilets jaunes.*

ABSTRACT

In this contribution we propose to contribute to the evaluation of algorithms called "word embedding" to the sociological analysis of texts: on the one hand, by comparing the results of semantic analyses of these algorithms with the now well-known approaches of textual data analysis; on the other hand, by focusing on what constitutes one of the main obstacles to the sociological analysis of the web: the difficulty to sociologically characterize the authors of statements from the web. To do this, we analyze the statements coming from two platforms of "civic tech" – the governmental platform, the "Grand Débat National", and its political and algorithmic response

1. EDF R&D, philippe.suignard@edf.fr
2. EDF R&D, caroline.escoffier@edf.fr
3. EDF R&D, lou.charaudeau@edf.fr
4. EDF R&D, mathieu.brugidou@edf.fr

proposed by a collective of Yellow Vests, the “Vrai Débat”. A third corpus from the “Entendre la France” platform, with the same design as that of the “Grand Débat National” and documented in terms of socio-political properties, will allow us to characterize the speakers according to their discourse and to try to predict, using machine learning approaches, the “pseudo properties” assigned to the speakers of the “Grand Débat National”.

Keywords: *word embedding, textual data analysis, civic tech, public debate, Yellow Jackets.*

1. Introduction

Dans cet article, nous nous proposons d'analyser les énoncés issus de deux plateformes de « civic tech » (Benvegnu, 2011 ; Mabi, 2014) – la plateforme gouvernementale du « Grand Débat National » (GDN) et sa riposte politique et algorithmique proposée par un collectif de Gilets jaunes (GJ), le « Vrai Débat » (VD), en combinant et en mettant à l'épreuve deux familles d'algorithmes dédiées à l'analyse de textes. Nous nous proposons de mettre en œuvre, d'une part, des approches éprouvées en analyse des données textuelles (ADT) (Reinert, 1986) sous Iramuteq qui ont montré récemment leur intérêt pour l'analyse de très grand corpus (Sebbah *et al.*, 2019) et, d'autre part, des méthodes nouvelles issues du croisement des mondes de l'informatique, de l'intelligence artificielle (IA) et du traitement automatique des langues (Cointet et Parasie, 2018). Nous nous intéresserons plus particulièrement à des familles d'algorithmes basées sur des « plongements de mots » (Mikolov *et al.*, 2013) et des « plongements de documents » via des méthodes de transfer learning (Devlin *et al.*, 2018 et Martin *et al.*, 2019).

Nous chercherons à savoir dans quelle mesure il est possible d'identifier avec ces méthodes la présence spéculaire d'un discours mais aussi d'un public Gilets jaunes au cœur même du dispositif GDN.

Nous nous interrogerons notamment sur les solutions méthodologiques permettant de qualifier les propriétés sociales des locuteurs sur lesquels on n'a que peu d'information directe – ce qui constitue un obstacle de taille pour l'analyse sociologique de ces plateformes mais aussi pour l'analyse des discours recueillis sur le web.

Répondre à ces différentes questions implique de s'inscrire dans une discussion sur ce que « font » ces différentes familles d'algorithmes à l'analyse sociologique des textes. Il s'agit en effet de préciser dans quelle mesure ces données « massives » et ces méthodes numériques réinterrogent, déplacent mais aussi retrouvent certaines des propositions épistémologiques les plus classiques de l'enquête sociologique – du moins dans sa tradition durkheimienne. Le statut des locuteurs – leur position sociale – apparaît en effet comme central dans l'analyse, dans la mesure où l'on cherche à le reconstituer sous forme de variables probabilistes ou « pseudo-variables ». Toutefois ces variables ne sont pas destinées à *expliquer* – y compris au sens sociologique – les énoncés, mais elles participent d'un faisceau d'indices (notamment discursifs) qui permettent d'*interpréter* les corpus analysés.

2. Hypothèses et corpus

Pour tenter de sortir de la crise politique suscitée par le mouvement des « Gilets jaunes », le Président de la République a appelé à un « Grand Débat National » comportant des réunions publiques et une importante phase de débat numérique grâce à une plateforme de délibération en ligne (https://granddebat.fr/). Des représentants des « Gilets jaunes » ont par ailleurs répliqué en proposant une plateforme alternative intitulée le « Vrai Débat » (https://le-vrai-debat.fr/). Très rapidement, et devant l'importance des corpus recueillis, la question de l'analyse et de la synthèse des propositions ou des échanges recueillis sur ces plateformes s'est posée : le recours aux méthodes d'IA ou « Big Data » destinées à traiter de très grands corpus de données est apparu aux différents acteurs[5] comme la seule solution possible pour agréger, hiérarchiser et classer ces propositions tout en respectant les réquisits de la démocratie participative et/ou délibérative. La conception même du type de démocratie engagée par ces débats et leurs implémentations numériques apparaît en effet comme un enjeu de recherche.

Les attentes à l'égard de ces approches apparaissent peut-être disproportionnées compte tenu de ce qu'elles sont réellement capables de faire (Bolaert et Ollion, 2018) et relèveraient

5. https://www.lemonde.fr/pixels/article/2019/02/01/grand-debat-en-ligne-et-democratie-l-analyse-et-la-transparence-des-donnees-en-question_5417911_4408996.html

ainsi assez classiquement d'une sociologie de la promesse (Vinck, 2015). Outre l'enjeu politique d'une mise en discussion de ce que les algorithmes font à la démocratie (Cardon, 2015), il y a bien un enjeu scientifique à tenter de cerner ce que peuvent faire (et ce que ne peuvent pas faire) ces méthodes d'analyse automatique. Cette discussion a notamment été ouverte par un article de Cointet et Parasie (2018) qui met en évidence ce que les différentes approches peuvent apporter à l'analyse sociologique des textes. Ils soulignent notamment l'intérêt d'une famille d'algorithmes dit de « plongement de mots » pour analyser de très grands corpus. Nous nous proposons de prolonger cette discussion de deux manières : d'une part en confrontant les analyses de plongement de mots aux approches bien connues maintenant d'analyse de données textuelles ou de textométrie ; d'autre part, en s'intéressant à ce qui constitue un des principaux obstacles à l'analyse sociologique du web : la difficulté à caractériser sociologiquement les auteurs des énoncés issus du web[6]. Bien que les données – et notamment textuelles – issues du web soient en effet « massives » – ce qu'indique le terme de « Big Data » – elles manquent pourtant singulièrement « d'épaisseur » (Peraya et Luengo, 2019) puisqu'on n'arrive pas, ou mal, à caractériser les agents ou les locuteurs en termes de propriétés sociales. La solution consistant à renouveler radicalement l'épistémologie des sciences sociales au profit d'un paradigme à construire de la trace (Boullier, 2015) paraît à ce jour peu convaincante. Cet obstacle apparaît dirimant concernant l'analyse du Grand Débat National. Nous proposons ici une stratégie – à portée limitée toutefois – pour pallier ce défaut.

De nombreuses questions portent sur la sociologie des participants à ces débats : les premiers résultats des travaux en cours suggèrent une sociologie très différente de ces deux publics. L'enquête du laboratoire PACTE[7] (Guerra *et al.*, 2019) réalisée auprès de groupes de GJ sur Facebook décrit un public majoritairement constitué de travailleurs précaires[8], habitant en territoire rural ou périurbain et refusant majoritairement de se situer sur l'axe gauche-droite. Dans l'enquête CEVIPOF (2019) auprès des participants aux réunions (RIL) du GDN, la moitié des personnes interrogées sont des retraités (âge moyen 57 ans), 54% déclarent s'en « sortir plutôt facilement avec les revenus du ménage » et 62% sont diplômés du supérieur. On note aussi une surreprésentation des habitants des grandes villes – notamment celles ayant placé en tête Macron lors du 1er tour de l'élection présidentielle.

Ces enquêtes donnent des indications précieuses mais rien ne nous permet d'extrapoler les résultats aux publics des débats numériques. Il n'y a pas de données décrivant le profil des participants aux plateformes de débat, hormis le code postal pour le GDN qui s'est avéré pour l'essentiel exploitable. Cette information permet de mettre en évidence une surreprésentation de la participation, pour les thématiques autour de la transition environnementale, du Sud Est et des grandes villes (voir aussi Bennani *et al.*, 2019).

Par ailleurs nous avons travaillé sur un troisième corpus (voir le tableau 1) issu de la plateforme « Entendre la France » (EF) dont le but était « *de permettre au plus grand nombre de Français de s'exprimer de la manière la plus simple possible, et d'être entendus* ». Le site permettait de répondre aux questions du GD, directement sur le site Web ou via Messenger. Il s'agit d'un public nettement plus jeune que celui du GDN – c'est d'ailleurs l'objectif poursuivi par les promoteurs de la plateforme. À la différence du GDN, EF posait une série de questions socio-démographiques (code postal, commune, type de commune, sexe, âge, formation, profession, taille de l'organisation) mais aussi celle du soutien aux GJ. La moitié des participants à la plateforme environ ont répondu à celles-ci. Elles constitueront des informations précieuses pour tenter de reconstituer les profils sociopolitiques des participants au GDN.

6. Voir notamment : Boyadjian J. (2016), Analyser les opinions politiques sur Internet. Enjeux théoriques et défis méthodologiques, Paris, Dalloz.
7. https://www.pacte-grenoble.fr/programmes/grande-enquete-sur-le-mouvement-des-gilets-jaunes
8. 67% peuvent être considérés en « situation précaire », le double de la moyenne nationale.

3. Comparer les discours du GDN et VD avec des approches ADT et IA

Notre premier objectif consiste à comparer les discours recueillis sur les plateformes GDN et VD à propos de la transition environnementale. D'un point de vue méthodologique, nous avons adopté une première approche textométrique classique (spécificités puis classification descendante hiérarchique sous Iramuteq[9]) et une approche par plongement de mots (Word2Vec, Mikolov *et al.*, 2013).

Tableau 1 – *Taille des corpus consacrés au thème de la transition environnementale*

Taille	Vrai Débat[10] sans arguments	Vrai Débat et arguments	Grand Débat National	Entendre la France
Nombre de textes	2 599	6 373	87 552	39 430
Nombre de formes	17 707	22 380	78 829	34 582
Nombre d'occurrences	225 039	351 991	21 764 365	1 273 520

3.1 L'analyse de données textuelles (ou ADT)

L'analyse des spécificités des mots (Chi2 de liaison à la classe) fait apparaître une sur-représentation dans le GDN[11] :

- des termes se rapportant au *dérèglement climatique* et à ses effets au niveau *mondial* ainsi qu'à d'autres enjeux environnementaux (*biodiversité*, *pollution*, *fossile*…) ;
- d'un lexique propre au registre moral et déontique : les citoyens sont invités par les propositions à une prise de conscience des problèmes et à changer de comportement ;
- d'un lexique macro-*économique* avec le thème de la *croissance* (*investir*, *entreprise*…).

L'analyse des spécificités fait apparaître dans le VD une liste d'enjeux sensiblement différents :

- la question de la *vitesse* et des *radars* sur les *routes*[12] ;
- les questions liées à l'alimentation (*alimentaire, légume, paysan, étiquetage, PAC*…) ;
- la *souffrance animale* notamment dans *les abattoirs* ;
- les conditions de vie notamment économiques (*prix, euro, facture, gratuité, vendre*) ;
- une série d'enjeux controversés : dénonciation d'un *marché* et de la *spéculation*, privatisations jugées indues, controverses impliquant *EDF* (*Linky, Bure*).

Une classification descendante hiérarchique (Reinert, 1983) sur l'ensemble du corpus formé par le GDN et le VD permet d'éclairer les thématiques privilégiées par l'une ou l'autre plateforme de débat. Ses résultats se sont avérés congruents avec les indications données par l'analyse de spécificités.

Le corpus du GDN consacré aux thématiques environnementales étant environ 60 fois plus important que celui du VD[13], la structure thématique mise en évidence est clairement celle du GDN. La première coupure de la classification (voir la figure 1) oppose une vision macroscopique de la transition écologique à une vision plus microscopique, centrée plutôt sur les pratiques. La branche de la classification plus « politique » abrite, d'une part, deux classes d'énoncés focalisés sur la biodiversité et l'agriculture et, d'autre part, cinq classes dont deux sur le climat et l'énergie, et trois soit à tonalité universaliste, soit discutant les conditions politiques ou économiques de

9. Pour une description plus complète de l'analyse textométrique, voir Brugidou *et al* (2020).
10. Corpus disponible sur https://www.le-vrai-debat.fr/syntheses/
11. Par ailleurs, une analyse des spécificités comparant l'ensemble des thématiques du GDN et du VD montre l'importance du lexique de la transition environnementale pour le GDN, le VD se caractérisant plus par les thématiques portant notamment sur la réforme de la démocratie (referendum, constitution…).
12. La limitation de vitesse à 80 km/h est un des éléments déclencheurs du mouvement, mais ses propositions ne portent ni seulement, ni en priorité sur cet enjeu (Zancarini et Ventresque, 2020).
13. Sur le problème de comparaison de corpus de tailles très différentes, voir Loubère et Marchand (2020).

la transition environnementale. Toutes ces classes, à l'exception de la classe sur le changement climatique, s'avèrent caractéristiques du VD, *i.e.* sont caractérisées par la variable illustrative VD. La branche portant sur les pratiques réelles ou jugées souhaitables comprend deux sous-groupes d'énoncés : le premier porte sur le transport, le second sur des pratiques liées au foyer. Toutes ces classes, à l'exception de la thématique du transport domicile/travail, sont caractéristiques du GDN.

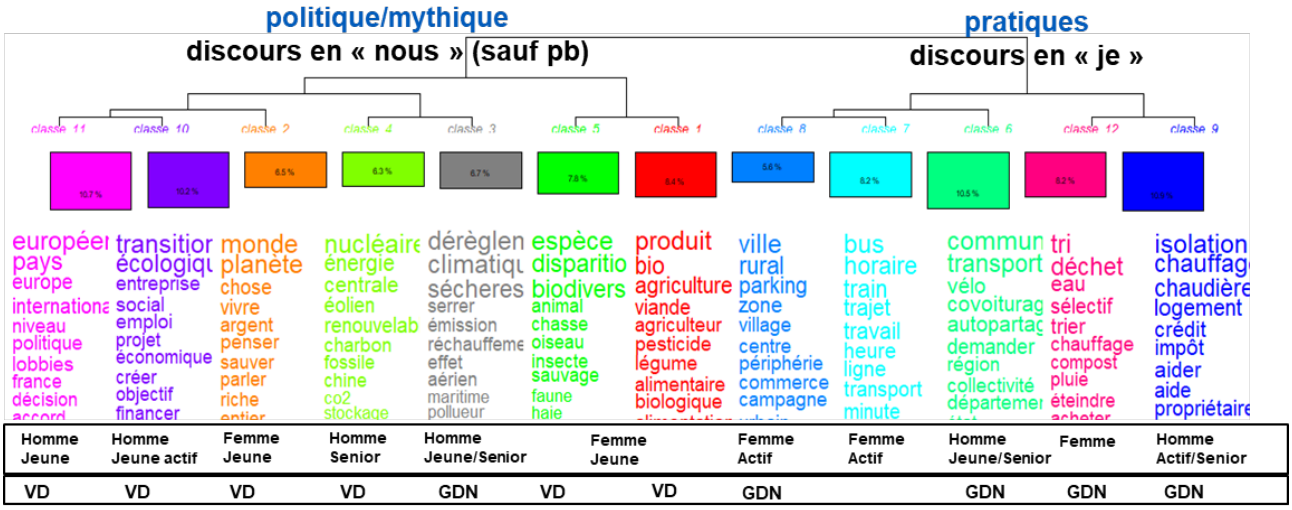

Figure 1 – *Dendrogramme de la classification descendante hiérarchique de l'ensemble des deux corpus de débat, pseudo-variables et spécificités (cf. §4)*

3.2 IA : plonger dans le corpus du GDN pour trouver la trace d'un discours « Gilets jaunes »

Notre deuxième approche consiste, non pas à comparer les vocabulaires du GDN et du VD pour identifier des différences, mais plutôt à repérer dans le corpus du GDN le vocabulaire caractéristique du VD. Autrement dit, après avoir reconnu les différences entre GDN et VD, nous cherchons à savoir si dans les contributions du Grand Débat National, on peut mettre en évidence un discours proche d'un discours « Gilets jaunes », *i.e.* qui présenterait des caractéristiques lexicales et thématiques des locuteurs du VD. La démarche utilisée est la suivante :

1. Des « embeddings » de mots sont calculés sur le corpus du GDN et sur celui du VD. La méthode utilisée est Word2Vec (Mikolov *et al.*, 2013). Elle consiste à transformer les mots sous la forme de vecteurs, avec pour idée générale que les mots ayant des contextes similaires auront des représentations vectorielles similaires[14].
2. Les représentations vectorielles de Word2Vec permettent des sommes, des soustractions ou des moyennes : les barycentres des corpus sont les moyennes des mots de chaque corpus.
3. Un calcul de spécificité est effectué à partir des similarités entre chaque mot du corpus et son barycentre : on cherche les mots les plus similaires au barycentre d'un corpus tout en étant les plus éloignés du barycentre de l'autre. Ce nouveau calcul permet de produire un indicateur de spécificité des mots, indicateur qui sera utilisé pour colorer les mots dans la figure 2 : bleue pour les mots très spécifiques au GDN et jaunes pour ceux spécifiques au VD.
4. Un fichier au format Gephi (Bastian *et al.*, 2009) est créé en transformant chaque mot en un point, un lien étant apposé entre deux mots si la similarité entre eux est supérieure

14. Un modèle est appris à partir de la concaténation des deux corpus GDN et VD avec les paramètres suivants : une fenêtre de 2 mots à gauche et 2 mots à droite, une couche cachée de taille 200, une architecture skip-gram, une fréquence minimale de 5 et 1000 itérations.

à un seuil fixé. La taille des points-documents est spécifiée par un calcul de PageRank (Page *et al*., 1999).

5. L'algorithme de spatialisation utilisé dans Gephi, de type « force-ressort », place ensemble les points/mots les plus similaires entre eux, e.g. *milliers*, *centaine* et *dizaine*.

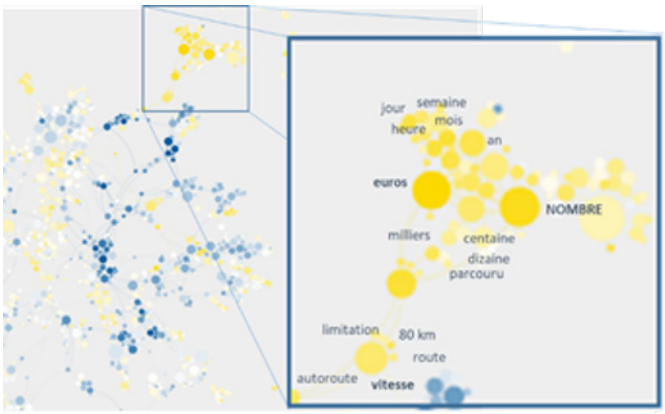

Figure 2 – *Zoom sur les mots du VD dans le GDN*

Au moins deux différences avec l'approche des spécificités précédente doivent être notées : d'abord, les mots utilisés seulement dans le VD n'apparaîtront pas dans cette analyse alors même qu'ils sont très caractéristiques de celui-ci. Ensuite, l'approche par plongement de mots capture le niveau du lexique mais aussi des variables que l'on peut qualifier de latentes ici comme la construction syntaxique, des formes ou motifs rhétoriques, etc.

Cette exploration nous met en présence de plusieurs « amas » lexicaux. On remarque ainsi le champ lexical des villes (*Lyon, Paris, Marseille, Bordeaux, Toulouse*), marque de l'ancrage territorial du discours et des mobilisations GJ. D'autres champs lexicaux se réfèrent à la thématique de la route et des limitations de vitesse (*limitation, vitesse, route, autoroute, 80 km*), associée au lexique des *nombres* et périodes de temps (*heure, jour, semaine, mois, ans*… très proche du premier ; cf. figure 2, mais aussi au carburant (*essence, éthanol, hydrogène, hybride*), au transport (*parking, péage, train, vélo, RER, métro, trajet, travail, domicile*), aux taxes (*taxe, TVA, surtaxé, écotaxe*) et plus largement au pouvoir d'achat que ce soit sous l'angle des prix (*prix, coût, tarif, montant*), des revenus (*revenu, salaire*) ou des prêts (*crédits, aide*). On note enfin des champs lexicaux se rapportant à l'alimentation (*céréales, lait, légumes, fruits, végétarien*) et plus particulièrement aux *cantines scolaires*.

Il ne s'agit ici que d'aperçus fugaces, d'indices ; il conviendrait bien sûr, dans une approche plus systématique, de désintriquer patiemment ces paquets de mots, en dépliant un à un les verbatim et les énoncés qui s'y trouvent condensés.

Par contraste, l'exploration par plongement de mots fait apparaître dans le GDN un discours (lexique et thématique) absent ou moins présent dans le VD. On note ainsi un champ lexical se rapportant à la pollution *chimique* (*pesticide, herbicide, dangereux*…) ou générée par les transports (*cargos, bateaux, kérosène*…). La désignation de *pollueurs* (*industrie, multinationale*) semble caractéristique du GDN ainsi que des *mesures* à mettre en œuvre (*actions concrètes*…) pouvant être contraignantes (*sanctions, lourdement, drastiques, contraignant, stopper, abandonner, interdire*…) ou incitatives (*promouvoir, encourager, développer*…). Par ailleurs, un important champ lexical relève du registre moral visant aux changements de *comportements* individuels des *citoyens* (*habitudes, modifier, changer, revoir, éduquer, informer, responsabiliser, sensibiliser, apprendre, école, gestes*).

Enfin, la topographie est très différente de celle valorisée dans le VD : les villes ne sont pas mises en avant mais des niveaux géographiques plus abstraits (*international, mondial, européen, local*) et des pays (*Inde, Usa, États-Unis*). Le GDN, plus que le VD, met ainsi en scène un contexte international pour traiter de la transition environnementale.

En guise de synthèse, on peut souligner que les deux approches « spécificités » et « plongement de mots » donnent des résultats qui présentent un air de famille – notamment le registre moral –, même si les listes de termes diffèrent en grande partie. Elles semblent ainsi manifester la présence d'un discours GJ tel qu'il apparaît dans le VD – attestée par des thématiques, un lexique voire un registre plus politique. À ce stade, très indiciel, disons qu'il existe des indices convainquant de cette présence alors même que les cadrages du GDN se révèlent particulièrement structurants. Il reste maintenant à savoir s'il est possible d'identifier un public Gilets jaunes dans le GDN et comment procéder pour cela.

4. Caractériser les locuteurs : un public GJ au cœur du Grand Débat ?

Pour caractériser les locuteurs, nous avons utilisé le corpus du site « Entendre la France »[15]. En plus des questions du GD, ils pouvaient renseigner les caractéristiques suivantes : code postal, commune, type de commune, sexe, âge, formation, profession, taille de l'organisation et position vis-à-vis des GJ. Nous avons retenu les variables suivantes :

– Sexe : 2 catégories : homme/femme16 ;
– Âge : 7 catégories réorganisées en 4 : « jeune », « jeune actif », « actif » et « senior »17 ;
– Position vis-à-vis des GJ : 3 catégories réorganisées en 2 : « soutient/ne soutient pas »18.

L'objectif était d'utiliser le contenu textuel des réponses pour prédire ces variables à l'aide de différentes techniques de « *Machine Learning* ».

4.1 Machine learning

Pour prédire chacune des trois variables, trois couples « données d'apprentissage/données de test » ont été constitués[19], données réparties en 70% pour l'apprentissage et 30% pour le test. Les méthodes suivantes ont ensuite été testées[20] : bayésien naïf, régression logistique, Word Embedding + Docov (Torki, 2018) et Bert et CamemBERT[21]. Le tableau 2 présente les résultats obtenus avec les 4 méthodes pour la prédiction du soutien aux Gilets jaunes, de la classe d'âge et du sexe. Les 4 méthodes fournissent sensiblement les mêmes résultats. Pour le « soutien », BERT est un peu en dessous[22] et a tendance à affecter très majoritairement les documents dans la catégorie « Soutien », ce qui explique son faible score de rappel. Seules les trois premières méthodes sont conservées dans la suite.

15. https://www.entendrelafrance.fr/
16. 5016 hommes et 3318 femmes
17. 4305 jeunes (18-24 ans), 1534 jeunes-actifs (25-34 ans), 944 actifs (35-54 ans), 713 séniors (55 ans et plus)
18. 3155 « Soutient » et 2361 « Ne soutient pas »
19. N'ont été gardées que les réponses des personnes ayant renseigné leur âge pour prédire la variable « âge ».
20. Nous décrivons ces méthodes dans (Brugidou *et al.*, 2020).
21. CamemBERT est une version de BERT entrainée sur des données françaises. L'hyperparamétrage pour le finetunig des modèles a été obtenu par grid search : la meilleure version a été entrainée avec un learning rate de 3e-5, sur 3 epochs et un warmup de 0.1.
22. La faiblesse des résultats obtenus avec BERT et CamemBERT (mêmes scores) est un peu surprenante : ces méthodes obtiennent de très bons résultats dans la plupart des tâches de classification. Cette faiblesse peut s'expliquer notamment par la longueur des textes à classifier et le peu de données d'apprentissage.

Tableau 2 – *Comparaison des 4 méthodes pour les 3 variables : soutien (*), classe d'âge (**) et sexe (***)*

	Naive Bayes			Regression			Docov			BERT
	*	**	***	*	**	***	*	**	***	*
Précision	0,640	0,422	0,665	0,613	0,393	0,648	0,631	0,361	0,642	0,680
Rappel	0,641	0,389	0,672	0,614	0,387	0,653	0,634	0,372	0,635	0,609
F-Mesure	0,640	0,400	0,659	0,613	0,389	0,638	0,631	0,347	0,598	0,589

D'après le tableau 2, on constate qu'il est plus facile de prédire le sexe des personnes, puis leur soutien aux GJ. Par contre l'âge est plus difficile à prédire car il y 4 classes à prédire et qu'elles sont déséquilibrées en nombre.

4.2 Enrichir l'analyse des données textuelles en documentant les propriétés des locuteurs

Le corpus « Entendre la France » nous a permis d'entrainer des classifieurs à prédire l'âge des répondants, leur sexe et leur soutien à la cause des GJ. Nous avons appliqué les classifieurs sur le corpus du GDN, puis un système de vote nous a permis de conserver le vote majoritaire entre ces 3 classifieurs et de calculer 3 nouvelles variables étoilées ajoutées au fichier Iramuteq.

Il est désormais possible d'attribuer ces propriétés sociopolitiques reconstituées aux locuteurs du GDN. Bien sûr, cette attribution est hypothétique : elle suppose notamment de considérer que le public de la plateforme Entendre la France (EF) présente les mêmes propriétés sociolinguistiques que les locuteurs s'exprimant sur la plateforme du GDN. Or, nous avons toutes raisons de croire que ces deux publics diffèrent : la plateforme d'EF a été créée pour pallier une participation supposée insuffisante des plus jeunes.

Tableaux 3 et 4 – *Composition sociopolitique du corpus « Entendre la France »[23] et de la prédiction appliquée au GDN - 87552 documents (en %)*

Entendre la France

soutient	56,3	jeune	62,3
ne soutient pas	43,7	jeune actif	19,3
homme	61,9	actif	10,5
femme	38,1	senior	7,9

Prédiction appliquée au GDN

soutient	36,0	jeune	39,0
ne soutient pas	64,0	jeune actif	40,8
homme	76,0	actif	16,5
femme	24,0	senior	3,7

Le profil des locuteurs d'EF est en effet beaucoup plus jeune que celui des participants aux réunions publiques (enquête CEVIPOF). La reconstitution de l'âge des participants du GDN double ainsi le nombre de jeunes actifs, (mais divise par 2 celui des seniors). Il est possible que le profil des seniors déclarés d'EF soit assez différent de celui des seniors du GDN. La prédiction de la variable de genre nous révèle par ailleurs un public du GDN sensiblement plus masculin que celui des participants à la plateforme EF. Enfin, l'algorithme nous donne une proportion de 36% de participants de la plateforme GDN soutenant le mouvement des GJ et de 64% ne le soutenant pas. Cette proportion apparaît vraisemblable : on s'attend en effet à une proportion forte de personnes ne soutenant pas les GJ dans le GDN lancé par le Président Macron et critiqué par ailleurs par les GJ. Toutefois, bien que minoritaire, elle est loin d'être négligeable – ce que nous laissait soupçonner une analyse même superficielle des discours issus du GDN. Cette proportion est par ailleurs comparable – environ 40% de soutien aux GJ – à celle trouvée par B. Monnery (2020) à partir d'une méthode différente (un modèle de régression construit à

23. % calculés sur le nombre de réponses qualifiées (âge : 10 637 ; sexe : 12 398 ; soutien : 6 278).

partir de l'analyse des réponses aux questions fermées sur EF pour expliquer le soutien aux GJ).

4.3 Des « pseudo-variables » interprétables ?

Nous disposons de deux méthodes pour mieux connaître les propriétés sociales des locuteurs du GDN, une première « indirecte » mais certaine caractérise le contexte sociodémographique du locuteur à partir de son lieu d'habitation via les codes postaux. La seconde emprunte une voie probabiliste mais qualifie directement le locuteur selon la stratégie exposée plus haut.

4.3.1 Approche indirecte et contextuelle.

La première méthode peut être qualifiée de « classique » dans la mesure où elle part des codes postaux[24] donnés par les participants au GDN pour territorialiser les classes de discours identifiées par classification descendante hiérarchique.

Trois des classes politiques, notamment sur les conditions politiques de la transition environnementale et sur l'énergie – caractéristiques du VD – sont le fait de locuteurs vivant en milieu urbain (centre-ville, grandes métropoles). Les classes d'énoncés sur l'agriculture, les pesticides et la biodiversité mais aussi sur le monde et la planète sont plutôt le fait de locuteurs vivant en milieu rural. La classe d'énoncés sur le changement climatique s'avère quant à elle caractéristique du périurbain aisé.

Du côté des classes portant sur les pratiques, l'analyse montre que les énoncés portant sur le transport domicile-travail mais aussi sur les transports en commun sont caractéristiques des locuteurs habitants des milieux urbains (grandes métropoles, banlieues) ou périurbain aisé. Les classes portant sur les transports urbain-campagne, l'isolation/chauffage et tris des déchets sont caractéristiques des locuteurs habitant en zone rurale (pour le transport), mixte ou périurbain.

4.3.2 Approche directe et probabiliste

L'approche par réseaux de neurones pose, on le sait, d'importants problèmes d'explicabilité (notamment Bolaert et Ollion, 2018). À la différence des approches de type régression, il est impossible de construire un modèle et de donner une valeur à différentes variables explicatives, tout simplement parce que nous n'en disposons pas. Mais, dans le cas qui nous occupe, il est possible de vérifier si des « pseudo-variables », ou encore des « propriétés probabilistes » donnent lieu à des interprétations intéressantes du point de vue de la sociologie politique. La figure 1 reproduit la classification analysée plus haut en caractérisant les classes d'énoncés avec les pseudo-variables issues du machine learning.

La plupart des classes d'énoncés privilégiant un registre politique et sur-employant la première personne du pluriel (voir la figure 1), que nous savons caractéristiques du VD, sont aussi caractéristiques des locuteurs qui soutiendraient les GJ selon l'algorithme d'apprentissage. Seule la classe sur le changement climatique, caractéristique du GDN, serait le fait de locuteurs soutenant les GJ. Inversement toutes les classes d'énoncés portant sur les pratiques seraient caractéristiques de locuteurs ne soutenant pas les GJ. Ces résultats tendent à conforter la qualité des pseudo-variables – au moins celle portant sur le soutien aux GJ.
L'analyse de la variable de genre et d'âge donne par ailleurs des résultats que l'on peut qualifier de « vraisemblables », ou de non contre-intuitifs : les femmes seraient ainsi surreprésentées parmi les locuteurs de la classe de discours universaliste (*monde/planète*), de la classe d'énoncés sur les produits bio et la biodiversité. Elles seraient aussi caractéristiques des classes transport rural-urbain et domicile-travail mais aussi tri des déchets. Les hommes seraient surreprésentés parmi les locuteurs des classes conditions politiques et économiques de la

24. Nous avons utilisé une typologie sur les codes postaux et les informations socioéconomiques INSEE disponibles, qui distingue notamment les zones urbaines périurbaines et rurales selon leur niveau de richesse.

transition environnementale, énergie, transport en commun et isolation/chauffage. La pseudo-variable sur les classes d'âge est sans doute moins facilement interprétable de manière isolée : elle vient le plus souvent préciser le profil des locuteurs des classes de discours, par exemple, les hommes actifs ou âgés préoccupés par l'isolation (ce qui implique le plus souvent d'être propriétaire).

5. Conclusions

A l'issue de ce parcours, plusieurs conclusions peuvent être tirées et sans doute autant de questions posées. Ces conclusions portent notamment sur les algorithmes et leurs usages, sur la sociologie des participants et leurs discours, et enfin contribuent à une discussion sur les fondements épistémologiques d'une analyse sociologique des textes.

Ces travaux posent une série de questions sur la comparaison entre des approches de type ADT et IA. Sans les détailler, on soulignera que les algorithmes de plongement de mots s'avèrent des outils d'exploration à la fois puissants et fins des thématisations par l'identification de champs lexicaux. D'autres recherches montrent, par ailleurs, les vertus des approches Reinert pour hiérarchiser des cadrages, et celles plus générales de la textométrie qui propose de calculer le poids des variables lexicales ou des formes de discours (par exemple les marques de l'argumentation) (Brugidou *et al.*, 2020).

Concernant la sociologie des publics, nous avons proposé une approche par *machine learning* qui permet d'inférer certaines des propriétés socio-politiques des locuteurs du GDN. Cette analyse sur le fond montre sinon la présence de locuteurs GJ, du moins de soutien à ce mouvement : l'analyse par le contexte territorial montre en effet la présence de locuteurs habitant des grandes agglomérations développant des thématiques par ailleurs caractéristiques du VD. La question de la sociologie des Gilets jaunes et de leurs soutiens reste très largement ouverte, on sait par ailleurs qu'elle a probablement varié dans le temps, ce qui plaiderait pour une analyse diachronique du corpus.

La question de la reconstitution de données absentes (pseudo-variables reconstituant les propriétés sociales et politiques des locuteurs) est à peine explorée : la plus grande prudence est de mise quant aux conditions de ces inférences. À défaut d'explicabilité des approches de *machine learning*, il nous semble possible de discuter de leur interprétabilité à la lumière notamment des travaux en sociologie politique sur le mouvement des GJ et sur le GDN.

Enfin, ces travaux ne sont pas sans conséquence sur l'épistémologie d'une analyse sociologique des textes. Ce type d'approche, par pseudo-variables, laisse en effet ouverte la question de savoir si ce sont les statuts des locuteurs qui déterminent les prises de position ou si ce sont ces dernières (les énoncés) qui, par leurs cheminements répétés, leurs chevauchements – certes cadrés par le design des plateformes de délibération et décrits par les algorithmes de plongement de mots – dessinent et renforcent ce qui finit par apparaître comme des structures positionnelles. Cette incertitude épistémologique rejoint selon nous une proposition d'une partie de la sociologie pragmatique[25], permettant de maintenir une interrogation sur les structures sociales, notamment comme le fruit d'accomplissements pratiques (ici discursives), sans pour autant ni se replier sur des positions déterministes, ni opter pour un insaisissable paradigme de la trace.

25. Il s'agit notamment d'articuler la perspective pragmatique avec des approches dispositionnelles et structurales, celles-ci étant décrites comme le produit d'accomplissements pratiques et non comme des causes, le fruit de processus et de sites d'agrégation qui les rendent descriptibles (Barthe *et al.*, 2013).

Références

Barthe Y., D. de Blic, J.-Ph. Heurtin, E. Lagneau, C. Lemieux et al. (2013), « Sociologie pragmatique : mode d'emploi », *Politix*, vol. 103, n° 3, pp. 175-204.

Bastian M., S. Heymann, and M. Jacomy (2009), « Gephi: an open source software for exploring and manipulating networks », *in* Third international AAAI conference on weblogs and social media.

Benvegnu N. (2011), « La politique des *netroots*. La démocratie à l'épreuve d'outils informatiques de débat public », Thèse de doctorat, CSI, École des Mines.

Bolaert J. et E. Ollion (2018), « The Great Regression – Machine Learning, Econometrics, and the Future of Quantitative Social Sciences », *Revue Française de Sociologie*, vol. 59, n° 3, pp.475-508.

Brugidou M., P. Suignard, C. Escoffier et L. Charaudeau (2020), « Un discours et un public « Gilets jaunes » au cœur du Grand Débat National ? Combinaison des approches IA et textométriques pour l'analyse de discours des plateformes « Grand Débat Nationa » et « Vrai débat » », à paraître dans les Actes des 15èmes JIADT.

Boullier D. (2015), « Les sciences sociales face aux traces du Big Data : société, opinion ou vibrations ? », *Revue française de science politique*, vol. 65, n° 5-6, pp. 805-828.

Cardon D. (2015), *À quoi rêvent les algorithmes. Nos vies à l'heure des big data*, Paris, Seuil, La République des idées.

Cointet J.-Ph. et S. Parisie (2018), « Ce que le *big data* fait à l'analyse sociologique des textes. Un panorama critique des recherches contemporaines », *Revue française de sociologie*, vol. 59, n° 3, pp. 533-557.

Devlin J., M.-W. Chang, K. Lee, and Kr. Toutanova (2018), « Bert: Pre-training of deep bidirectional transformers for language understanding », eprint arXiv:1810.04805.

Guerra T., Chl. Alexandre and F. Gonthier (2020), « Populist Attitudes Among the French Yellow Vests », *Populism*, vol. 3, n° 1.

Loubère L. et P. Marchand (2020), « Les thématiques du Grand Débat et du Vrai Débat : approche textométrique », Journées d'études, Sciences Po Paris, 16 et 17 janvier 2020.

Mabi C. (2014), « Le débat CNDP et ses publics à l'épreuve du numérique : entre espoirs d'inclusion et contournement de la critique sociale », Thèse en science de l'information et de la communication, sous la direction de Laurence Monnoyer-Smith et de Serge Bouchardon.

Martin L., B. Müller, P. J. Ortiz Suarez, Y. Dupont, L. Romary, E. Villemonte de la Clergerie, D. Seddah, and B. Sagot (2019), « CamemBERT: a Tasty French Language Model », eprint arXiv:1911.03894.

Mikolov T., I. Sutskever, Ch. Kai, Gr. Corrado, and D. Jeffrey (2013), « Distributed representations of words and phrases and their compositionality », *in* Curran Associates Inc. (ed.), *Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2*, pp. 3111-3119.

Monnery B. (2020), « Qui a participé au GrandDébat.fr ? Prédiction du soutien aux Gilets jaunes chez les contributeurs à partir d'une deuxième plateforme », Journées d'études, Sciences Po Paris, 16 et 17 janvier 2020.

Page L., S. Brin, R. Motwani, and T.Winograd (1999), « The pagerank citation ranking: Bringing order to the web », Technical report, Stanford InfoLab.

Peraya D. et V. Luengo (2019), « Les *Learning Analytics* vus par Vanda Luengo », *Distances et médiations des savoirs*, vol. 27. URL : http://journals.openedition.org/dms/4096

Reinert M. (1983), « Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte », *Cahiers de l'Analyse des Données*, vol. 8, n° 2, pp. 187-198.

Reinert M. (1986), « Un logiciel d'analyse lexicale », *Cahiers de l'Analyse des Données*, vol. 11, n° 4, pp. 471-481.

Sebbah B., N. Souillard, L. Thiong-Kay et N. Smyrnaios (2018), « Les gilets jaunes, des cadrages médiatiques aux paroles citoyennes », Rapport de recherche préliminaire - 26 novembre 2018, Laboratoire d'Études et de Recherches Appliquées en Sciences Sociales, Axe Médias et médiations socio-numériques – Université de Toulouse : https://www.lerass.com/wp-content/uploads/2018/01/Rapport-Gilets-Jaunes-1.pdf.

Torki M. (2018), « A document descriptor using covariance of word vectors », *in* Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers), pp. 527-532.

Vinck D. (2015), « Les *digital humanities* comme promesse pour et par les sciences humaines », in M. Audétat (éd.), *Sciences et technologies émergentes : pourquoi tant de promesses ?*, Paris, Éditions Hermann, pp. 131-145.