



Pour plus de transparence dans l'analyse automatique des consultations ouvertes : leçons de la synthèse du Grand Débat National

Aurélien Bellet, Pascal Denis, Rémi Gilleron, Mikaela Keller, Nathalie Vauquier

► To cite this version:

Aurélien Bellet, Pascal Denis, Rémi Gilleron, Mikaela Keller, Nathalie Vauquier. Pour plus de transparence dans l'analyse automatique des consultations ouvertes : leçons de la synthèse du Grand Débat National. 2020. hal-02860659v2

HAL Id: hal-02860659

<https://hal.inria.fr/hal-02860659v2>

Preprint submitted on 23 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pour plus de transparence dans l'analyse automatique des consultations ouvertes : leçons de la synthèse du Grand Débat National

Aurélien Bellet, Pascal Denis, Rémi Gilleron, Mikaela Keller, Nathalie Vauquier
Université de Lille, Inria Lille & CRIStAL (UMR CNRS 9189)

23 octobre 2020

Résumé

Face aux limites de la démocratie représentative, les consultations numériques participatives publiques permettent de solliciter, à différents niveaux de pouvoir, des contributions de citoyens pour essayer de mieux impliquer les individus dans les décisions politiques. Leur conception et leur mise en place posent des problèmes bien connus tels que les biais dans les questions ou la représentativité de la population participante. Nous considérons, dans cet article, les problèmes nouveaux liés à l'utilisation de méthodes issues de l'intelligence artificielle pour l'analyse automatique de contributions en langage naturel. Réaliser une telle analyse est un problème difficile pour lequel il existe de nombreuses méthodes reposant sur des hypothèses et des modèles variés. En considérant comme cas d'étude les contributions aux questions ouvertes du Grand Débat National¹, nous montrons qu'il est impossible de reproduire les résultats de l'analyse officielle commandée par le gouvernement. En outre, nous identifions des choix arbitraires non explicités dans l'analyse officielle qui conduisent à émettre des doutes sur certains de ses résultats. Nous montrons également que différentes méthodes peuvent mener à des conclusions différentes. Notre étude² met ainsi en lumière la nécessité d'une plus grande transparence dans les analyses automatiques de consultations ouvertes pour assurer leur reproductibilité et la confiance du public dans leur restitution. Nous concluons par des pistes d'amélioration des consultations participatives et de leur analyse pour qu'elles puissent inciter à la participation et être des outils utiles au débat public.

1 Introduction

La question de la représentation des citoyens possède une longue histoire en politique [15]. Les limites de la démocratie représentative contemporaine, qui peine à rendre compte des opinions des citoyens et de leur diversité, nourrissent un intérêt renouvelé pour des modes de représentation participatifs permettant d'impliquer plus directement et plus rapidement

1. <https://granddebat.fr/>

2. Les données et le code source sont disponibles sur <https://gitlab.inria.fr/magnet/gdn-lib>.

les individus dans les prises de décisions politiques [5, 2]. Ces dernières années ont ainsi vu l'émergence de plates-formes numériques participatives, qui servent en outre des objectifs d'e-gouvernance³ [22]. De telles consultations ont été organisées dans plusieurs pays, à différents niveaux de pouvoir, et dans des domaines aussi stratégiques que l'urbanisme, l'aménagement du territoire, l'environnement, la santé, l'éthique, et l'économie⁴. En théorie, l'avantage est double, puisque ce type de démarche pourrait mener à des prises de décisions potentiellement mieux informées et aussi ressenties comme plus justes, car intégrant un plus grand nombre de personnes et de points de vue.

La mise en place de ce type de consultations participatives pose cependant des défis importants. En premier lieu, il s'agit de s'assurer de l'absence de biais dans le choix et la formulation des questions. En second lieu, ces consultations étant ouvertes à un large public et s'appuyant souvent sur les technologies de l'information, il convient d'assurer une bonne représentativité des contributeurs ou, à défaut, de prendre en compte les spécificités de la population répondante dans l'analyse et les résultats.

Au-delà de ces premiers défis relativement bien étudiés dans la littérature [5], nous considérons dans cet article *le problème des possibles biais dans l'analyse automatique des réponses*. En effet, les données issues de ces consultations numériques sont typiquement de grande taille et ne peuvent être traitées exhaustivement par des humains. L'analyse doit donc être réalisée par des traitements automatiques qu'il est essentiel de valider. Les consultations étaient jusque récemment limitées à des questionnaires fermés, c'est-à-dire avec des choix de réponses encadrés, avec des méthodes d'analyse essentiellement basées sur des statistiques élémentaires et donc facilement vérifiables. Mais des techniques récentes issues de l'Intelligence Artificielle (IA), en particulier le Traitement Automatique des Langues (TAL) et l'Apprentissage Artificiel (AA), offrent désormais la possibilité d'analyser — voire de comprendre — automatiquement des textes générés librement par les citoyens⁵. Ceci permet a priori de s'affranchir des questionnaires fermés et de proposer des questionnaires ouverts, c'est-à-dire des questions auxquelles le participant a toute liberté de répondre en texte libre. Néanmoins, ce nouveau type de questionnaire renouvelle les problématiques relatives à l'analyse des réponses. En effet, l'analyse automatique de textes libres est une tâche complexe, eu égard à la quantité et à la nature des données ainsi qu'à leur structuration syntaxique et sémantique, et dont l'évaluation est difficile. Si la prise en compte politique est primordiale, il est essentiel, a minima, d'assurer une forme de confiance du public dans la restitution des consultations. Pour cela, il est nécessaire de définir des bonnes pratiques que nous tenterons d'ébaucher à partir d'une étude critique d'un exemple concret de consultation ouverte.

Dans le présent article, nous nous focalisons sur le Grand Débat National (GDN), la

3. http://portal.unesco.org/ci/fr/ev.php-URL_ID=3038&URL_DO=DO_TOPIC&URL_SECTION=201.html. “L'e-gouvernance est l'utilisation par le secteur public des technologies de l'information et de la communication dans le but d'améliorer la fourniture d'information et de service, d'encourager la participation du citoyen au processus de décision et de rendre le gouvernement plus responsable, transparent et efficace.”

4. Voir, par exemple, ce rapport sur les initiatives européennes : <https://www.vie-publique.fr/en-bref/270833-cour-des-comptes-europeenne-evalue-les-consultations-publiques>

5. Voir [7] pour des consultations locales issues de la plate-forme québécoise Votepour.ca.

consultation citoyenne initiée en janvier 2019 par le président de la République française. Conçu en réponse au mouvement des Gilets Jaunes, qui fit rage dans les mois précédents, le GDN avait pour objectif affiché de « redonner la parole aux Français sur l’élaboration des politiques publiques qui les concernent »⁶. À côté d’autres phases de consultations organisées directement sur le terrain, un questionnaire en ligne, fait de questions ouvertes et fermées, fut lancé entre janvier et mars 2019. Ce questionnaire, divisé en quatre thématiques indépendantes, contient de nombreuses questions ouvertes ayant chacune suscité plusieurs dizaines de milliers de réponses textuelles parfois longues et argumentées. L’analyse de cette consultation en ligne a été confiée à OPINIONWAY⁷ qui a produit des documents de synthèse : un sur la méthodologie et quatre sur les résultats des analyses par thème. Comme indiqué précédemment, nous ne considérons pas ici les problèmes de biais dans les questions, de représentativité de la population et de biais dans l’analyse statistique des questions fermées et renvoyons sur ces sujets aux communiqués des organisations de contrôle⁸ et aux études récemment publiées [8, 3].

Nous nous concentrons sur l’analyse automatique des contributions textuelles correspondant aux réponses aux questions ouvertes. Les auteurs de l’analyse officielle ont choisi, pour chaque question ouverte, de déterminer des catégories et des sous-catégories définies par des intitulés textuels courts et de fournir comme résultat des taux de répartition dans ces catégories (un exemple est présenté en Figure 1). L’analyse a été réalisée en utilisant des méthodes automatiques d’IA propres à l’entreprise avec des interventions humaines. Malheureusement, l’ensemble de la chaîne de traitement mise en place manque cruellement de transparence, malgré les demandes [13] et les annonces⁹ faites en ce sens. Ainsi, les différentes étapes d’analyse, autant automatique qu’humaine, restent inconnues. Or, il existe de nombreuses méthodes de catégorisation de données textuelles pouvant mener à des choix de catégories différents et à des répartitions différentes, différences qui peuvent être accentuées par les traitements humains réalisés au cours de l’analyse. De plus, les affectations individuelles des réponses aux catégories n’ont pas été rendues publiques. Cette opacité sur les méthodes utilisées, le code et les résultats empêche toute validation externe, et a fortiori toute reproduction de l’analyse officielle réalisée.

Contributions et résultats. À défaut de pouvoir reproduire la méthode d’analyse et de pouvoir la valider sur les assignations des réponses aux catégories, nous proposons dans cet article une rétro-analyse qui nuance les conclusions de l’analyse officielle et met en lumière le besoin de transparence. Plus spécifiquement, en nous appuyant sur différentes méthodes de l’état de l’art en TAL, décrites ici exhaustivement, nous proposons différentes approches

6. <https://www.nouvelobs.com/politique/20181219.0BS7366/grand-debat-national-voulu-par-macron-a-quoi-faut-il-s-attendre.html>

7. www.opinion-way.com

8. <https://www.participation-et-democratie.fr/communiqu-e-de-l-observatoire-des-debats>

9. Le secrétaire d’État chargé du numérique a déclaré sur Twitter le 8 avril 2019 : “L’ensemble des données et des méthodologies utilisées seront rendues publiques. Certains logiciels ne sont pas en open source mais il importe d’être le plus précis possible sur le paramétrage choisi afin que les chercheurs puissent faire leur travail de vérification.” Voir https://twitter.com/cedric_o/status/1115171491384037377

pour (ré-)affecter les textes des contributions aux catégories et sous-catégories choisies dans la synthèse officielle. En plus d'assurer le caractère reproductible de nos expériences, cette méthodologie a pour avantage d'illustrer le fait que différents algorithmes, voire différents paramétrages du même algorithme, donnent des répartitions différentes. Mais surtout, aucune des approches mises en œuvre ne permet d'obtenir des résultats proches de ceux de l'analyse officielle. Notre étude permet en outre d'identifier des questions difficiles à analyser, qui ont typiquement généré des réponses longues et argumentées. L'étude de ces questions nous amène à questionner la validité de certains résultats et de certains choix de l'analyse officielle. Grâce à des annotations manuelles, nous avons par exemple pu mettre en évidence une sous-estimation des taux donnés dans la synthèse officielle pour certaines catégories, et identifié des catégories non répertoriées qui représentent pourtant davantage de contributions que certaines catégories présentes dans la synthèse officielle.

Notre étude démontre que l'analyse par des méthodes de TAL et d'IA est susceptible de biais à cause de nombreux choix sur la représentation des données textuelles, sur la méthode d'analyse et son paramétrage, et sur les éventuelles interventions humaines dans les traitements. Ce problème bien connu en sociologie statistique est exacerbé par l'utilisation de méthodes d'IA et interroge le rôle de l'analyste quand il s'agit de traiter des questions politiques. Conformément à un courant actuel en intelligence artificielle [11] et dans toutes les sciences expérimentales, il est donc essentiel d'assurer la reproductibilité d'une analyse en diffusant les données et les codes. À défaut de pouvoir publier le code, les résultats fournis doivent permettre une validation extérieure de l'analyse et donner l'opportunité aux participants de savoir comment leur contribution a été comptabilisée. Nous pensons, par ailleurs, qu'il est important de confronter des analyses obtenues par différentes approches si l'on veut réduire les biais et obtenir une synthèse fiable. Ces préconisations sont particulièrement essentielles dans le cas de contributions participatives à visée politique.

En complément de cette nécessité de transparence, notre étude suggère des pistes d'amélioration pour les consultations participatives futures et leur analyse. En particulier, l'ouverture des résultats permet d'envisager des consultations plus collaboratives et interactives dans lesquelles les participants ont l'opportunité d'intervenir sur l'utilisation de leurs contributions, par exemple en corrigeant les affectations à une catégorie. Ceci fournirait une supervision humaine partielle qui pourrait être utilisée pour améliorer l'analyse initiale.

Plan de l'article. La Section 2 présente les données du GDN et la synthèse officielle, puis discute des limites de cette dernière. La Section 3 présente notre méthodologie et les résultats principaux de notre rétro-analyse (les résultats détaillés se trouvent en annexe). Nous concluons par une discussion sur les bonnes pratiques et les pistes d'amélioration en Section 4.

2 Les questions ouvertes et les résultats de l'analyse officielle

Cette section décrit de manière plus détaillée les questions ouvertes du questionnaire en ligne du GDN, les réponses qui ont été fournies par les contributeurs et la synthèse produite

Thème	Questions	Contributions			Catégories de la synthèse	
	nombre	nombre min	nombre max	moyenne	nombre min	nombre max
FDP	8	39 431	154 150	119 000	5	24
ORG	25	6 047	84 452	33 000	2	24
DC	31	13 326	89 910	67 000	3	18
TE	14	611	128 751	74 000	3	18

TABLE 1 – Statistiques sur les questions ouvertes par thème, les nombres minimum, maximum et moyen des contributions (ou réponses) correspondantes, et les effectifs minimum et maximum des catégories définies par la synthèse officielle.

par OPINIONWAY et QWAM.

2.1 Description des questions ouvertes et de leurs réponses

Pour rappel, la consultation du GDN lancée le 21 janvier 2019 et clôturée le 18 mars 2019 fut organisée selon 4 thèmes, à savoir :

- Fiscalité et Dépenses Publiques (FDP)
- Organisation de l’Etat et des Services Publics (ORG)
- Démocratie et Citoyenneté (DC)
- Transition Ecologique (TE)

Le questionnaire en ligne contient au total 52 questions fermées à choix multiples et 78 questions ouvertes. Pour chacun des thèmes, le nombre de contributions est de l’ordre de 500 000 et le nombre de participants de l’ordre de 400 000 avec un taux variable de réponses selon les questions.

Nous nous intéressons dans cet article aux questions ouvertes. Une question est dite *ouverte* dès lors que les participants peuvent y répondre sous forme d’un texte libre, et ne sont donc pas limités à un nombre prédéfini de choix, par exemple : « Que faudrait-il faire pour mieux représenter les différentes sensibilités politiques ? ». Certaines de ces questions ouvertes sont dites *filtrées* lorsqu’elles spécifient une question préalable, typiquement une question polaire (c’est-à-dire en « oui-non »), pour laquelle elle suppose une réponse particulière. Par exemple, « Diriez-vous que votre vie quotidienne est aujourd’hui touchée par le changement climatique ? », suivi de la question filtrée « Si oui, de quelle manière votre vie quotidienne est-elle touchée par le changement climatique ? ». Il y a, sur l’ensemble des 78 questions des quatre thèmes, 22 questions filtrées. Ces dernières seront traitées comme les questions ouvertes en prenant les contributions des personnes ayant répondu à la réponse attendue à la question du filtre (« oui », sur l’exemple précédent) même s’il existe parfois des réponses aux questions filtrées après une réponse négative au filtre.

Quelques statistiques élémentaires sur ces questions ouvertes, ainsi que sur les contributions (c’est-à-dire, les réponses à ces questions) sont reprises dans la Table 1. On constate que le nombre de contributions est très variable selon les thèmes, avec notamment des nombres moyens de réponses allant de 33 000 pour le thème ORG à 119 000 pour le thème FDP.

Une variabilité importante est aussi présente à l'intérieur de chaque thème, les écarts les plus amples étant au sein du thème TE. En ce qui concerne la longueur des questions et des contributions, mesurées en nombre de mots, notons tout d'abord que les questions ont une taille moyenne de 17.3 mots tout thème confondu. La taille des contributions, quant à elle, varie aussi très fortement selon les questions. Ainsi, la longueur moyenne des réponses (non vides) varie entre 8 et 130 mots, la médiane varie entre 2 et 66, et le troisième quartile varie entre 3 et 178. Cette dernière valeur signifie que, sur une question, un quart des contributions ont une longueur supérieure à 178 mots.

Ces statistiques élémentaires montrent une participation importante, avec de nombreux participants ayant exprimé des opinions détaillées et argumentées dans leurs réponses. Indépendamment d'une prise en compte politique des réponses, ceci suggère que l'analyse devrait contenir un retour à destination des participants pour leur montrer comment leurs réponses individuelles ont été prises en compte dans les résultats.

2.2 Le format de la synthèse officielle

L'analyse de cette consultation en ligne a été confiée à l'entreprise OPINIONWAY, spécialisée en sondages politiques et études marketing, à la demande du gouvernement¹⁰. La synthèse officielle¹¹, produite en avril 2019, contient une notice méthodologique et quatre synthèses, une par thème. Chacune des synthèses contient des statistiques descriptives sur les répondants (répartition temporelle, géographique et territoriale) et deux volets, correspondant à des sous-ensembles de questions distincts. Le premier volet consiste en un ensemble de statistiques compilées par OPINIONWAY sur les réponses aux 52 questions fermées. Les possibles biais ici sont liés au choix des questions et aux spécificités de la population répondante plutôt qu'à l'analyse, qui consiste simplement en un comptage des réponses.

Nous nous focalisons sur le second volet qui est une synthèse de l'analyse des réponses aux questions ouvertes. Cette analyse a été déléguée par OPINIONWAY à l'entreprise QWAM¹², spécialisée dans l'analyse des contenus (en particulier textuels). On retrouve, comme dans le cas des questions fermées, le problème de biais dans les questions et le problème de représentativité. Nous ne traitons pas ces problèmes mais remarquons que la liberté procurée aux répondants par les questions ouvertes leur permet de relever certains problèmes avec des réponses comme « *C'est de la foutaise, toutes les questions sont orientées!!! On est pas là pour répondre à un QCM!"* ». Nous nous concentrons sur l'analyse automatique de ces 5 millions de réponses rédigées par les participants aux questions ouvertes.

Les auteurs de l'analyse officielle ont choisi de livrer une synthèse qui prend la forme, pour chaque question ouverte, d'un ensemble de catégories avec des pourcentages de répartition des réponses dans celles-ci (voir l'exemple de la Figure 1 que nous utiliserons comme illustration). Chaque catégorie (entre 2 et 24 selon les questions, cf. Table 1) est définie

10. Suite à un marché public organisé par le Service d'information du gouvernement (SIG) remontant à 2015, voir https://www.liberation.fr/france/2019/02/15/grand-debat-des-algorithmes-et-de-l-ia-pour-trier-classer-et-sous-classer-les-idees_1709678

11. <https://granddebat.fr/pages/syntheses-du-grand-debat>

12. <http://www.qwamci.com/>

Question 5. Que faudrait-il faire pour mieux représenter les différentes sensibilités politiques ?
 Question ouverte - Réponses non suggérées - Plusieurs réponses possibles

118 356
contributions

Introduire la proportionnelle	42,0%
Instaurer une dose de proportionnelle	40,4%
Instaurer la proportionnelle intégrale	1,7%
Autres contributions	9,9%
Autres éléments sur les partis politiques	2,5%
Autres éléments sur la démocratie	1,7%
Autres éléments sur les élus	1,1%
Autres éléments sur les députés	0,9%
Autres éléments sur l'élection présidentielle	0,8%
Autres éléments sur le respect	0,8%
Autres éléments sur les associations	0,5%
Autres éléments sur les formations	0,5%
Modifier les règles des scrutins	8,0%
Prendre en compte le vote blanc	2,6%
Réformer les élections législatives	1,9%
Conserver le scrutin majoritaire	1,1%
Instaurer le vote obligatoire	0,9%
Réglementer le temps de parole pendant les campagnes	0,5%
L'importance d'une majorité stable	0,4%
Réformer le Parlement	3,7%
Améliorer la représentativité de l'Assemblée nationale	2,6%
Réformer ou supprimer le Sénat	1,6%
Renforcer la démocratie directe	1,9%
Des référendums	1,0%
Le Référendum d'Initiative Citoyenne (RIC)	0,7%
Impliquer la société civile	0,4%
Rien	0,6%
Autres contributions trop peu citées ou inclassables	20,0%
Non réponses	31,5%

FIGURE 1 – Un exemple de résultat de l'analyse d'une question ouverte : la question 5 du domaine « La démocratie et la citoyenneté ».

par un intitulé textuel. Certaines de ces catégories se divisent en sous-catégories, également définies par un contenu textuel et dont le nombre total peut aller jusque la centaine pour certaines questions. Le cas de la Figure 1 fait apparaître 9 catégories et 20 sous-catégories. Par exemple, l'une des catégories est dénommée « Modifier les règles de scrutin » et possède des sous-catégories « Prendre en compte le vote blanc », . . . , « L'importance d'une majorité stable ».

Pour chacune des questions, il existe aussi deux catégories particulières : les « non-réponses », la catégorie qui comptabilise les textes vides, et les « inclassables », la catégorie des réponses n'ayant pu être rangées dans aucune des catégories choisies par l'analyse officielle. En règle générale, les intitulés des catégories et sous-catégories prennent la forme d'une réponse à la question en jeu ; ainsi, « Modifier les règles des scrutins » ou « Réformer le Parlement » sont des réponses sémantiquement valides à la question « Que faudrait-il faire pour mieux représenter les différentes sensibilités politiques ? ». Il y a cependant des exceptions avec des choix plus discutables comme la catégorie « Autres contributions » et toutes ses sous-catégories ou encore la sous-catégorie « L'importance d'une majorité stable ». On peut aussi noter, toujours sur la question de la Figure 1, la difficulté à voir une différence sémantique entre les catégories « Autres contributions » et « inclassables ». La longueur moyenne des intitulés de catégories est de 5.5 mots, alors les intitulés des sous-catégories tendent à être un peu plus riches avec 7.1 mots en moyenne.

Lorsqu'une catégorie possède des sous-catégories, les répartitions sont également données pour les sous-catégories. Si nous poursuivons l'exemple de la Figure 1, nous pouvons constater que la somme des taux des sous-catégories de la catégorie « Réformer le parlement » est supérieure au taux de la catégorie alors que la somme des taux des sous-catégories de la catégorie « Renforcer la démocratie directe » est inférieure au taux de la catégorie. Cet exemple nous permet de constater que les catégories peuvent se chevaucher (une réponse peut être affectée à plusieurs catégories) : il y a $100 - 31,5 - 20 = 49,5\%$ de réponses affectées à des catégories choisies (en ignorant les non-réponses et les « inclassables ») mais la somme des taux de répartition dans les catégories choisies est de $42 + 9,9 + 8 + 3,7 + 1,9 + 0,4 + 0,6 = 66,5\%$. Ces constats sont valables pour l'ensemble des questions ouvertes. Nous pouvons donc affirmer que, pour chacune des questions ouvertes, une réponse d'une catégorie peut être classée dans une seule sous-catégorie, plusieurs sous-catégories, voire aucune sous-catégorie.

En résumé, les catégories et sous-catégories sont définies par un intitulé textuel, qui prend généralement la forme d'une réponse à la question posée. Les réponses non vides sont classées dans des catégories et une réponse peut être affectée à plusieurs catégories. Les catégories peuvent se diviser en sous-catégories et une réponse de la catégorie peut être classée dans une seule, plusieurs, voire aucune sous-catégorie(s). Certaines réponses ne correspondant à aucune des catégories choisies sont rangées dans une catégorie particulière, celle des inclassables. L'analyse officielle est donc une catégorisation (création de groupes de réponses) hiérarchique (avec des catégories découpées en sous-catégories) avec recouvrements (les sous-catégories ne définissent pas une partition mais peuvent se recouvrir).

Avant de nous intéresser à la méthode employée pour obtenir cette synthèse, on peut d'ores-et-déjà émettre certains constats critiques. Le choix des (intitulés des) catégories et sous-catégories est parfois discutable car sans lien avec la question posée. De plus, nous avons noté que les catégories peuvent se recouvrir mais on peut constater que les taux de recouvrement de la synthèse officielle sont faibles au vu des intitulés des catégories. Enfin, nous notons que le taux d'inclassables pour l'ensemble des questions est très élevé (entre 15 et 30% selon les questions). Ceci signifie que, pour une question avec 100 000 contributions, entre 15 000 et 30 000 d'entre elles ne sont pas prises en compte dans la synthèse officielle. Or, nous avons vu que les participants avaient fait l'effort de rédiger des réponses parfois longues aux questions. Ces réponses dites « inclassables » peuvent correspondre à des « coups de gueule » ou à des idées originales. Ceci montre que l'objectif de catégorisation thématique aurait pu être complété par d'autres axes comme l'analyse des sentiments (ex : réponses exprimant de la colère) ou la détection d'idées émergentes (ex : une réponse supportant une énergie nouvelle particulière comme l'hydrogène).

2.3 Réflexions sur la méthode de l'analyse officielle

Nous avons vu que le problème choisi dans l'analyse officielle, sur base des questions ouvertes et de leurs réponses, consiste à déterminer des catégories en vue de répartir les réponses au sein de ces catégories. Plus précisément, il s'agit donc, pour chaque question ouverte et les réponses textuelles associées, de :

1. Déterminer des catégories et sous-catégories sémantiquement pertinentes ;
2. Affecter les réponses à ces catégories et sous-catégories ;
3. Calculer les pourcentages de répartition.

Vu sous l’angle de l’apprentissage artificiel, ce problème relève de la *classification non supervisée* (ou *clustering*)¹³. Elle est *non supervisée* dans la mesure où on ne dispose pas a priori des catégories et donc a fortiori pas d’exemples d’affectation de réponses dans les catégories. Le problème considéré ici est rendu encore plus complexe par le fait que le nombre des catégories n’est pas non plus connu a priori et que ces catégories ont une structure hiérarchique, puisque certaines catégories donnent lieu à des sous-catégories, et qu’elles peuvent se recouvrir. L’étiquetage des catégories (et sous-catégories) par un intitulé textuel est également un défi supplémentaire, puisque celui-ci doit avoir une sémantique pertinente relativement à la question posée. En outre, la forme des catégories est sujette à des contraintes supplémentaires, de concision notamment, pour apparaître dans un compte-rendu lisible par le plus grand nombre. Ce problème de catégorisation est rendu d’autant plus difficile par la nature textuelle des données. À ceci s’ajoute le format particulier du question-réponse, les différents domaines et registres de langue mobilisés (du très technique au très général, du très formel au très oral), le recours à l’humour et l’ironie, ainsi qu’à d’autres formes d’intention amenant des contributeurs à éviter de donner des réponses informatives.

De manière générale, les problèmes d’apprentissage non supervisés sont connus comme difficiles en apprentissage artificiel, de surcroît quand il s’agit d’analyser des corpus textuels. Ils sont l’objet de recherches très actives à l’intersection des communautés TAL et AA. Les méthodes applicables à la catégorisation de données textuelles sont nombreuses, et aucune méthode n’est connue comme meilleure pour toutes les tâches. Ainsi, les méthodes traditionnelles de clustering “plat” (c’est-à-dire non hiérarchique) telles que les k -moyennes ou le clustering spectral, leurs extensions hiérarchiques, ainsi que les algorithmes de classification supervisée (comme Bayes naïf, régression logistique, les SVM ou plus récemment les réseaux de neurones profonds), éventuellement combinés avec des méthodes de réduction de dimension, sont couramment utilisés (voir [16] pour un survol). Plus spécifiquement orienté vers la catégorisation avec recouvrement de données textuelles, l’algorithme Latent Dirichlet Allocation (LDA) [4] existe en version plate ou hiérarchique et contient une réduction de dimension implicite à travers la modélisation par variables latentes des sujets (*topics*) abordés dans les textes. L’ensemble de ces méthodes suppose un codage numérique des textes : l’approche dominante consiste à les représenter sous la forme de vecteurs (nous reviendrons sur ce point dans la Section 3, et de manière plus détaillée dans l’Appendice A).

Parmi ce large choix de techniques et de paramétrages associés, quelle est donc la méthode retenue pour obtenir la synthèse officielle ? Selon le rapport d’étude¹⁴, QWAM a utilisé des méthodes internes qui sont « des algorithmes puissants d’analyse automatique des données textuelles en masse (big data), faisant appel aux technologies du traitement automatique du langage naturel couplées à des techniques d’intelligence artificielle (apprentissage

13. Le lecteur peut se reporter au Chapitre 16 de [16] pour une introduction au clustering de textes.

14. <https://granddebat.fr/media/default/0001/01/f73f9c2f64a8cf0b6efa24fdc80179e7426b8cc9.pdf>

profond/deep learning) ». Les résultats issus de cette analyse automatique ont ensuite été post-traités par des humains : « une intervention humaine systématique de la part des équipes qualifiées de QWAM et d’OPINIONWAY pour contrôler la cohérence des résultats et s’assurer de la pertinence des données produites ». À notre connaissance, il s’agit là des seules informations disponibles publiquement sur l’approche utilisée. Nous pouvons faire les constats suivants :

- Les codes des algorithmes ne sont pas fournis et ne sont pas ouverts.
- La méthode de choix des catégories, des sous-catégories, de leur nombre et des intitulés textuels associés n’est pas spécifiée.
- La méthode d’affectation des textes aux catégories n’est pas non plus spécifiée.
- Les affectations individuelles des réponses aux catégories ne sont pas fournies.
- Malgré l’intervention humaine avérée, aucune mesure d’évaluation des catégories par des humains (par exemple des scores d’accord entre annotateurs) n’est fournie pour attester de la cohérence des groupes de réponses trouvés relativement à leur sémantique.

Nous ne pouvons donc que conjecturer que l’analyse officielle a été réalisée en combinant des algorithmes de classification non supervisée de textes, avec une supervision humaine ou un post-traitement humain pour affiner la catégorisation et attribuer une sémantique textuelle aux groupes de réponses. La part d’intervention humaine est difficile à saisir, mais on suppose très certainement une approche conjointe de recherche de mots-clés dans les textes d’un groupe et une expertise humaine. Il est par ailleurs probable que l’analyse ait été effectuée de manière itérative, avec plusieurs étapes de catégorisation et de définition des intitulés textuels avant d’arriver aux résultats définitifs présentés dans la synthèse officielle.

En résumé, l’approche utilisée par l’entreprise QWAM pour construire la synthèse officielle est parfaitement opaque. Nous ne disposons même pas d’élément permettant de retrouver à quelle catégorie est affectée une réponse particulière dans l’analyse officielle. Il est ainsi impossible de vérifier si une autre approche fournit des affectations identiques ou similaires. Nous avons donc du travailler uniquement à partir des taux de répartition dans les catégories fournis dans la synthèse officielle, seul et maigre matériau rendu publiquement disponible.

3 Ré-affecter les réponses aux catégories de l’analyse officielle

À défaut de pouvoir reproduire l’approche mise en place par l’analyse officielle, nous proposons une forme de rétro-analyse. En prenant comme point de départ les catégories choisies dans la synthèse officielle, nous utilisons différentes méthodes d’affectation des réponses dans les catégories basées sur l’état de l’art et montrons que ces méthodes peuvent produire des résultats différents. Nous montrons la difficulté, voire l’impossibilité, à retrouver les taux de répartitions de la synthèse officielle. Notre étude nous permettra également de mettre au jour certaines anomalies dans les résultats de l’analyse officielle.

3.1 Modélisation du problème

Nous commençons par décrire brièvement notre approche générale. Les données et le code source pour reproduire nos expériences sont mis à disposition publiquement.

3.1.1 Objectif

Nous supposons connues les catégories et sous-catégories de l'analyse officielle avec leurs intitulés textuels. Par souci de clarté, nous nous limitons à l'affectation des réponses dans les catégories (et ignorons donc les sous-catégories). Les réponses vides sont rangées dans la catégorie « Non réponses ». Nous considérons alors le problème suivant :

1. pour chaque question ouverte,
2. avec en entrée l'ensemble des réponses textuelles non vides à la question,
3. affecter les réponses aux catégories.

Une réponse peut être affectée à une, plusieurs, ou aucune catégorie. Une réponse qui n'est affectée dans aucune catégorie sera alors affectée à la catégorie particulière des inclassables. Ceci permet de construire des répartitions avec les mêmes contraintes que celles de la synthèse officielle.

3.1.2 Approche

Considérons une question Q . Toute réponse est naturellement représentée par le contenu textuel R de la réponse. Pour pouvoir affecter une réponse à une catégorie, nous représentons également chaque catégorie par un contenu textuel comme suit. Nous représentons une catégorie par le texte C constitué de la concaténation de l'intitulé textuel de la catégorie et des intitulés textuels de toutes ses sous-catégories. Sur l'exemple de la Figure 1, la catégorie « Réformer le parlement » sera représentée par le texte « Réformer le parlement. Améliorer la représentativité de l'assemblée nationale. Réformer ou supprimer le sénat ».

Nous pouvons alors, de façon très naturelle et bien éprouvée en TAL et en recherche d'information, considérer une fonction de distance (ou une similarité) entre textes et définir une méthode générale d'affectation des réponses aux catégories comme suit :

1. pour chaque réponse R et pour chaque catégorie C ,
2. calculer la distance entre la réponse R et la catégorie C ,
3. affecter la réponse R à la catégorie C si la distance est inférieure à un seuil.

Dans cette méthode, le seuil peut être choisi identique pour toutes les catégories ou une valeur peut être choisie pour chacune des catégories. Les réponses qui n'ont été affectées à aucune catégorie, c-à-d celles pour lesquelles la distance calculée est supérieure au seuil pour toutes les catégories, sont affectées à la catégorie des inclassables.

Nous avons ainsi remplacé un problème de catégorisation non supervisée par un problème plus simple d'affectation de réponses textuelles à des catégories elles-mêmes définies par un intitulé textuel. Après exécution de notre algorithme, nous trouvons des groupes de réponses

pour chacune des catégories et en déduisons des taux de répartition de réponses dans les catégories que nous pouvons comparer aux taux fournis par l’analyse officielle. Il nous reste à préciser la représentation des textes et la distance utilisée.

3.1.3 Choix de la distance et des représentations vectorielles du texte

Les représentations vectorielles de mots et de textes connaissent une histoire longue et mouvementée en TAL avec des développements récents, principalement avec les réseaux de neurones profonds qui ont mené à des avancées importantes sur de nombreuses tâches, dont la traduction automatique. Une description de ces travaux et des représentations choisies pour notre étude, ainsi qu’une discussion sur les limites de notre approche, peuvent être trouvées dans l’Appendice A. Nous utiliserons quatre représentations vectorielles, appelées **tf-idf**, **mangoes**, **fasttext** et **bert**, que nous pensons représentatives de l’état de l’art. Nous mesurerons la distance entre les représentations vectorielles des textes par la distance cosinus, dont les valeurs sont comprises entre 0 et 1,

3.2 Résultats

La méthode présentée en Section 3.1.2 nécessite de déterminer des seuils d’affectation des réponses textuelles aux catégories. Nos choix se basent sur la façon de retrouver les taux présentés dans l’analyse officielle, ce qui va nous mener à choisir un seuil unique ou un seuil spécifique par catégorie.

3.2.1 Seuil unique basé sur le taux d’inclassables

Dans cette première expérience, nous choisissons *un seuil unique pour chaque approche de représentation des textes*, déterminé comme suit. Rappelons que pour un seuil donné, une réponse est classée dans une catégorie si la distance entre la représentation vectorielle du texte de la réponse à la représentation vectorielle du texte de la catégorie est inférieure au seuil. Nous obtenons ainsi des taux de répartition des réponses dans les catégories choisies par l’analyse officielle et pouvons en déduire un taux d’inclassables (réponses affectées à aucune des catégories). *Le seuil unique est choisi de façon à obtenir le même taux d’inclassables que dans l’analyse officielle.*

Par exemple, pour la question 1 du thème « Démocratie et Citoyenneté » (« En qui faites-vous le plus confiance pour vous faire représenter dans la société et pourquoi? »), en choisissant le seuil pour avoir un taux d’inclassables de 13.2% comme dans la synthèse officielle, nous obtenons les répartitions présentées dans la Table 2.

Nous constatons que notre méthode a tendance à sur-évaluer les taux de répartition par rapport à la synthèse officielle et que nous trouvons des groupes de réponse avec un fort recouvrement, c-à-d que beaucoup de contributions sont affectées à plusieurs catégories. À la vue des intitulés des catégories, ces affectations multiples sont plausibles et il est probable, comme nous l’avons déjà signalé, qu’elles soient trop limitées dans la synthèse officielle. Par exemple, la catégorie « Les modalités du vote, des élections » est certainement sous-estimée

Catégorie	officiel	tfidf	mangoes	fasttext	bert
les élus locaux	27.4%	31.10%	38.67%	30.83%	32.37%
contributions sur le manque de confiance	23.2%	34.41%	37.40%	48.10%	49.84%
les élus	22.2%	27.64%	36.80%	34.62%	21.27%
les corps intermédiaires	8.7%	7.00%	7.96%	1.85%	5.00%
pourquoi	4.9%	6.05%	14.53%	27.09%	15.39%
les modalités du vote, des élections	3.8%	22.13%	22.41%	31.07%	24.46%
les conditions de confiance	2.6%	12.73%	16.08%	34.73%	31.96%
fait confiance sous réserve	1.6%	11.14%	6.13%	6.89%	15.68%
la société civile	1.4%	9.72%	4.36%	13.23%	31.64%
les maires et les députés	1.3%	9.89%	35.23%	19.45%	13.50%
les référendums	1.2%	0.65%	3.84%	0.21%	3.90%
autres contributions	0.6%	11.43%	2.16%	18.30%	17.17%
Inclassables	13.2%	13.2%	13.2%	13.2%	13.2%
Total	98.9%	171.2%	225.6%	266.4%	262.2%

TABLE 2 – Pourcentages de répartitions des réponses dans les catégories obtenus avec différentes approches pour la question « En qui faites-vous le plus confiance pour vous faire représenter dans la société et pourquoi ? ». Pour chaque approche, le seuil est choisi de manière à obtenir le même taux d’inclassables que dans l’analyse officielle (ici, 13.2%).

dans la synthèse officielle avec un taux de 3.8% bien inférieur au taux attendu au vu de la question. Avec notre méthode et quelle que soit la représentation choisie, nous obtenons des taux qui sont sans doute plus proches de la réalité.

La Table 2 révèle deux premiers résultats importants : d’une part, qu’il existe des différences de répartition importantes entre l’analyse officielle et notre approche, et d’autre part, que différentes méthodes de représentations des textes donnent des taux d’affectation différents. Ces deux conclusions, illustrées ici sur une question spécifique, sont validées sur l’ensemble des questions comme le montre l’étude statistique présentée dans l’Appendice B.

3.2.2 Seuils reproduisant les taux de chaque catégorie

Un seuil unique basé sur le taux d’inclassables ne permettant pas de retrouver les répartitions de la synthèse officielle, nous choisissons, dans cette deuxième expérience, un seuil par catégorie. Le seuil pour chacune des catégories est choisi de telle manière que la méthode présentée en Section 3.1.2 affecte un nombre de réponses correspondant à celui de la synthèse officielle. Les différentes catégorisations ne diffèrent alors que par l’identité des réponses affectées à chaque catégorie et non leur nombre. Nous souhaitons étudier si les affectations individuelles des réponses sont similaires ou pas selon la méthode choisie. Malheureusement, il nous est impossible de considérer l’analyse officielle car les affectations individuelles des réponses aux catégories ne sont pas connues. Nous menons donc cette étude en considérant notre méthode d’affectation dans les catégories avec nos différentes représentations vectorielles. Cette étude, présentée en Appendice C, montre que la même méthode utilisée avec des représentations différentes produit des affectations comparables pour des ca-

*tégories aux réponses particulièrement simples et des affectations très différentes pour des catégories avec des intitulés peu explicites et regroupant un grand nombre de réponses. On observe par ailleurs que les représentations textuelles plus riches telles que **bert** permettent de capturer davantage de sémantique.*

3.2.3 Anomalies dans les taux de répartition de la synthèse officielle

Dans l'expérience précédente, nous avons choisi les seuils pour retrouver les taux de répartition de l'analyse officielle. Nous avons réalisé une étude statistique de la répartition de ces seuils (voir Appendice B) qui met en évidence des questions et catégories avec des seuils particulièrement bas (proches de 0) ou élevés (proches de 1). Une étude spécifique de ces questions, présentée ci-après, nous permet d'identifier certaines anomalies dans les répartitions de la synthèse officielle.

Le cas des seuils très bas. On peut noter que tous les seuils calculés sur des catégories avec des textes courts (« non », « rien », « aucun ») sont très bas. Par exemple, pour la question 17 « Y a-t-il d'autres points sur la transition écologique sur lesquels vous souhaiteriez vous exprimer ? » du thème TE, la catégorie 14 « non » de cette question nécessite un seuil très faible de 0.046 pour obtenir le bon taux de réponses à savoir 1.7% de 153 809 réponses. Une étude qualitative des réponses classées dans la catégorie par notre méthode montre que les affectations sont correctes. En effet, les 2644 réponses classées « non » par notre méthode se répartissent en 2506 réponses « non », 137 réponses « non » avec un signe de ponctuation additionnel et une réponse mal affectée « non au glyphosphate ». Cependant, malgré le seuil bas, si on regarde les 50 réponses immédiatement supérieures au seuil, on trouve 42 nuances de « non » (« non merci », « pas pour le moment ») et 8 réponses diverses difficiles à affecter. Les réponses suivantes nécessiteraient une expertise humaine pour être classées. Il semble en tout cas que le nombre de réponses affectées à la catégorie est sous-estimé par la synthèse officielle à cause de la difficulté à capturer des variantes textuelles élaborées d'un simple « non ».

Le cas des seuils très élevés. Ceci se produit pour des questions et catégories contenant des réponses textuelles souvent longues de plusieurs mots à quelques phrases. Prenons l'exemple de la question 10d du thème TE « Et qui doit selon vous se charger de vous proposer ce type de solutions alternatives ? » et la catégorie « les acteurs publics ». Avec la représentation vectorielle **mangoes**, il faut un seuil particulièrement élevé de 0.838 pour trouver 67 369 réponses classées dans cette catégorie correspondant au pourcentage de 43.4% de l'analyse officielle. Une étude qualitative est plus difficile que pour le cas précédent car les réponses sont plutôt longues (25 000 réponses ont plus de 20 mots). Nous avons cependant *classé manuellement* les réponses dans la catégorie « les acteurs publics » et avons trouvé un taux de 54.5%. Ceci montre que la catégorie est largement sous-estimée par l'analyse officielle avec une différence de l'ordre de 15 000 réponses entre l'affectation manuelle et la synthèse officielle. Si on compare l'annotation manuelle avec les affectations fournies par

notre méthode avec la représentation **mangoes**, nous voyons apparaître dans les réponses affectées à la catégorie par notre méthode des faux positifs comme « les constructeurs », « les employeurs » et « les entreprises ». Nous avons également montré qu’il y a 72% de faux négatifs, c’est-à-dire 72% de réponses qui sont des acteurs publics au sens de l’affectation manuelle mais qui ne sont pas rangés dans cette catégorie par notre méthode à cause de la difficulté de classer des réponses textuelles longues.

Ceci met en évidence la difficulté de bien affecter les réponses dans les catégories et nous fait *émettre des doutes sur le bien-fondé des taux indiqués dans la synthèse officielle* qui, comme nous l’avons dit précédemment, ne fournit aucun élément d’évaluation pour juger de la pertinence des groupes de réponses associées aux catégories.

3.2.4 Anomalies dans le choix des catégories de la synthèse officielle

Enfin, une analyse qualitative de la question 10d du thème TE (« Et qui doit selon vous se charger de vous proposer ce type de solutions alternatives ? ») nous permet de relever certains biais dans le choix des catégories. En effet, une étude « manuelle » des réponses nous a fait remarquer que certaines réponses concernaient la prise en charge par l’individu lui-même des solutions alternatives. Aucune catégorie dans la synthèse officielle ne correspondant à ce type de réponse, nous avons donc souhaité étudier si le nombre de ces réponses était significatif pour savoir si cette catégorie aurait dû apparaître dans la synthèse.

Pour cela, nous avons considéré une catégorie « prise en charge par l’individu » et avons *classé manuellement* les réponses comme correspondant, ou pas, à cette catégorie. Nous avons trouvé un taux de 4.5% des réponses pour cette catégorie, soit environ 7000 réponses. Celles-ci ont des formes diverses pouvant être courtes comme « moi même » [*sic*], « les citoyens », « je suis grand », . . . , « c’est mon problème », ou plus longues comme « les français [*sic*] sont assez intelligents pour les trouver seuls », . . . , « les citoyens sont les premiers maîtres de leur choix ». Pour cette question, l’analyse officielle a choisi des catégories de plus faible effectif qui ne correspondent d’ailleurs pas toujours à des réponses à la question posée. Ceci met en évidence *un certain arbitraire dans le choix des catégories de la synthèse officielle*, qui ignore des réponses ayant une sémantique forte pour l’étude comme la nécessaire implication des individus dans les changements de comportement.

4 Conclusions et perspectives

Dans cet article, nous avons étudié un exemple de consultation participative à large échelle : le Grand Débat National. Mettant de côté les questions plus classiques sur le biais dans les énoncés des questions et sur la représentativité de la population répondante, nous avons souhaité étudier les questions soulevées par l’analyse des réponses textuelles à des questions ouvertes en utilisant des méthodes d’IA.

Nous avons tout d’abord présenté la synthèse officielle et relevé la complète **opacité de la méthode utilisée pour produire l’analyse**. Nous avons également soulevé plusieurs questionnements sur la synthèse, notamment quant au grand nombre de réponses délibéré-

ment exclues de la synthèse, au choix des intitulés des catégories et au faible recouvrement des réponses affectées à ces catégories.

Nous avons ensuite présenté une rétro-analyse de la synthèse officielle s'appuyant sur des méthodes de l'état de l'art en TAL pour ré-affecter les contributions textuelles aux catégories proposées par la synthèse officielle avec une méthodologie transparente et reproductible. Nous en tirons trois conclusions principales :

- Même en reprenant les mêmes catégories, **nous ne sommes pas parvenus à retrouver des effectifs comparables à ceux de la synthèse officielle**, quelles que soient la méthode de représentation vectorielle et la manière de mesurer les distances entre ces représentations. Comme évoqué précédemment, notre approche peut certainement être améliorée en prenant mieux en compte certaines spécificités du problème. Une partie des différences observées peut aussi s'expliquer par le manque de transparence sur la méthode et les résultats de la synthèse officielle, qui rend impossible des comparaisons plus fines. Mais dans tous les cas, les informations fournies ne permettent pas de valider la synthèse officielle.
- Comme illustré dans notre étude par les différents choix possibles de représentation vectorielle des textes, **différentes approches de l'état de l'art aboutissent à des résultats différents**. La synthèse officielle n'est ainsi qu'une interprétation possible du contenu des contributions, parmi de nombreuses autres. Du point de vue de l'utilisateur de ces technologies ou du commanditaire d'une étude, il convient donc de ne pas tirer des conclusions trop tranchées à partir des résultats d'une seule analyse.
- Notre étude permet de mettre en évidence **des problèmes dans la synthèse officielle**, en particulier une sous-estimation de certains effectifs et des biais sur le choix des catégories et sous-catégories. Sans remettre en cause le sérieux ni l'impartialité du travail des entreprises OPINIONWAY et QWAM, nos résultats montrent que la synthèse officielle est pour le moins imparfaite.

Ces conclusions nous amènent à formuler des suggestions de bonnes pratiques pour les analyses automatiques de consultations participatives futures. Si l'avenir de ces consultations dépend en premier lieu de leur prise en compte politique, il repose également sur des analyses capables de démontrer leur transparence et la prise en compte des contributions des participants. Nous proposons plusieurs pistes en ce sens :

- **Introduire davantage de transparence pour permettre la reproductibilité du traitement automatique des consultations participatives**. Les techniques utilisées doivent être clairement décrites (y compris dans leur paramétrage), avec idéalement une ouverture du code quand cela est possible. La chaîne de traitement dans son ensemble (comprenant le traitement humain) doit également être précisément définie. Enfin, il est nécessaire de publier les résultats obtenus à une granularité suffisamment fine pour permettre une validation indépendante (par des citoyens, des associations ou encore des chercheurs) avec un débat contradictoire, et pour que tout participant puisse connaître la façon dont ont été prise en compte sa contribution. Dans le cas

du GDN, cela aurait pu prendre la forme d'une publication des affectations de chaque contribution individuelle aux catégories.

- **Considérer différents axes d'analyse et confronter différentes méthodes d'analyse** dans les conditions de transparence indiquées ci-dessus. Dans le cas du GDN, la répartition des contributions dans des catégories thématiques aurait pu être complétée par la recherche de propositions émergentes (par exemple, une réponse supportant une énergie nouvelle particulière comme l'hydrogène) et d'émotions (par exemple, relever les réponses exprimant de la colère). Par ailleurs, pour un axe d'analyse donné, il existe plusieurs méthodes reposant sur des hypothèses et biais spécifiques, donnant donc une interprétation possible du corpus. La confrontation de plusieurs analyses est utile pour nuancer certaines conclusions et ainsi mener à une synthèse finale plus fiable.
- **Concevoir des consultations plus collaboratives et interactives** dans lesquelles les participants ont l'opportunité d'intervenir sur la manière dont leurs contributions sont utilisées. Il est notamment possible de s'appuyer sur les participants pour obtenir une supervision humaine partielle qui permet de déployer des méthodes d'apprentissage supervisé ou semi-supervisé plus fiables que les méthodes non supervisées, et dont les performances peuvent être évaluées de manière quantitative. Différentes manières de solliciter cette supervision humaine peuvent être considérées. Une première approche consiste à faire annoter des textes par des volontaires (voir l'initiative de la Grande Annotation¹⁵). Une seconde est de donner la possibilité aux participants de signaler des erreurs dans la manière dont leur contribution a été prise en compte, pour ensuite effectuer un nouvel apprentissage s'appuyant sur ces corrections. Une troisième option est de permettre aux participants de commenter ou de voter sur les contributions des autres et d'utiliser ces informations pour guider l'analyse. Quoi qu'il en soit, donner une véritable place aux citoyens dans le processus d'analyse a le potentiel de favoriser la participation, et donc la représentativité de la population répondantes.

Comme nous l'avons montré, réaliser une analyse de contributions ouvertes avec des méthodes d'IA est une tâche difficile. La production d'une telle synthèse ne doit donc pas relever d'un choix arbitraire et opaque d'un analyste se substituant au politique, mais doit au contraire être transparente, contradictoire, et donner un rôle de supervision aux participants. Les pistes ci-dessus permettent de favoriser l'implication des citoyens et leur acceptation du processus, qui constituent une condition indispensable à l'avenir de telles consultations et à leur utilité pour le débat public.

Références

- [1] Sanjeev ARORA, Yingyu LIANG et Tengyu MA : A simple but tough-to-beat baseline for sentence embeddings. *In International Conference on Learning Representations (ICLR)*, 2017.

15. <https://grandeannotation.fr/>

- [2] Marie-Hélène BACQUÉ et Yves SINTOMER : *La démocratie participative : Histoire et généalogie*. La Découverte, 2011.
- [3] Hamza BENNANI, Pauline GANDRÉ et Benjamin MONNERY : Les déterminants locaux de la participation numérique au grand débat national : une analyse économétrique. *Revue économique*, 71(4):715–737, 2020.
- [4] David M BLEI, Andrew Y NG et Michael I JORDAN : Latent dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022, 2003.
- [5] Loïc BLONDIAUX : *Le Nouvel Esprit de la démocratie. Actualité de la démocratie participative*. Seuil, 2008.
- [6] Piotr BOJANOWSKI, Edouard GRAVE, Armand JOULIN et Tomas MIKOLOV : Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics (ACL)*, 5:135–146, 2017.
- [7] Eya BOUKCHINA, Sehl MELLOULI et Emna MENIF : From citizens to decision-makers : A natural language processing approach in citizens’ participation. *International Journal of E-Planning Research (IJEPR)*, 7(2):20–34, 2018.
- [8] Éric BUGÉ et Camille MORIO : Le grand débat national, apports et limites pour la participation citoyenne. *Revue du droit public*, 5:1206–1239, 2019.
- [9] Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE et Kristina TOUTANOVA : BERT : pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.
- [10] Edouard GRAVE, Piotr BOJANOWSKI, Prakhar GUPTA, Armand JOULIN et Tomas MIKOLOV : Learning word vectors for 157 languages. In *International Conference on Language Resources and Evaluation (LREC)*, 2018.
- [11] Odd Erik GUNDERSEN et Sigbjørn KJENSMO : State of the art : Reproducibility in artificial intelligence. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1644–1651, 2018.
- [12] Matt KUSNER, Yu SUN, Nicholas KOLKIN et Kilian WEINBERGER : From word embeddings to document distances. In *International Conference on Machine Learning (ICML)*, pages 957–966, 2015.
- [13] LE CONSEIL SCIENTIFIQUE DE LA SOCIÉTÉ INFORMATIQUE DE FRANCE (SIF) et Aurélien BELLET : Grand débat et IA : quelle transparence pour les données? *Libération*, page 25, 8 avril 2019.
- [14] Omer LEVY et Yoav GOLDBERG : Neural word embedding as implicit matrix factorization. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2177–2185, 2014.
- [15] Bernard MANIN : *Principes du gouvernement représentatif*. Flammarion, 3ème édition, 2019.
- [16] Christopher D. MANNING, Prabhakar RAGHAVAN et Hinrich SCHÜTZE : *Introduction to Information Retrieval*. Cambridge University press, 2008.

- [17] Tomas MIKOLOV, Kai CHEN, Greg CORRADO et Jeffrey DEAN : Efficient estimation of word representations in vector space. *In International Conference on Learning Representations (ICLR)*, 2013.
- [18] Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg S CORRADO et Jeff DEAN : Distributed representations of words and phrases and their compositionality. *In Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, 2013.
- [19] Jeffrey PENNINGTON, Richard SOCHER et Christopher MANNING : Glove : Global vectors for word representation. *In Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.
- [20] Matthew E. PETERS, Mark NEUMANN, Mohit IYYER, Matt GARDNER, Christopher CLARK, Kenton LEE et Luke ZETTLEMOYER : Deep contextualized word representations. *In Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL-HLT)*, 2018.
- [21] Gerard SALTON, Anita WONG et Chung-Shu YANG : A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.
- [22] Anthony TROTTA : *Advances in E-Governance : Theory and Application of Technological Initiatives*. Productivity Press, 2017.
- [23] Peter D. TURNEY et Patrick PANTEL : From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, 37:141–188, 2010.

A Description des représentations vectorielles de textes

Dans cet appendice, nous présentons le principe général des représentations vectorielles de textes ainsi que les représentations choisies pour notre étude, et discutons des limites de notre approche.

A.1 Principe général

Les représentations vectorielles des mots et des textes possèdent une longue histoire en analyse automatique, remontant au moins aux travaux fondateurs de Salton [21] en recherche d’information. Les modèles vectoriels ont aussi largement imprégné les recherches en TAL, d’abord dans le cadre de la sémantique distributionnelle [23] puis plus récemment dans le contexte des *word embeddings* (ou *plongements lexicaux*) neuronaux [18].

Schématiquement, les premières représentations ont d’abord été définies par expertise selon la tâche considérée comme, notamment, la représentation TF-IDF (pour “term frequency - inverse document frequency”) en recherche d’information ou la représentation PPMI (pour “positive pointwise mutual information”) en sémantique distributionnelle. Dans ce type de représentations, chaque document (respectivement, mot) est représenté par un vecteur dont les composantes encodent les scores associés à la fréquence d’occurrence (respectivement, de co-occurrences) de certains termes d’un vocabulaire préalablement défini (plusieurs centaines de milliers, voire millions de mots). Étant données la grande taille des vocabulaires et les distributions très asymétriques liées aux fréquences de mots, ces vecteurs sont extrêmement grands et creux (ils contiennent beaucoup de zéros). Ce type de représentations a, par ailleurs, le défaut de ne pas être en mesure de capturer naturellement la proximité entre des mots pourtant proches sémantiquement (p.ex., *emploi* et *travail* ou *taxe* et *impôt*), puisque chaque paire de termes correspond à une composante différente du vecteur.

Ces limitations ont conduit les chercheurs en TAL et RI à investiguer des méthodes capables de générer des représentations plus denses et de plus faible dimension (typiquement entre 50 et 1000), à même de mieux modéliser ces proximités. Celles-ci se sont basées tout d’abord sur des méthodes classiques de réduction de dimensions et de factorisation de matrices, puis sur des méthodes par apprentissage neuronal depuis les années 2010, telles que GLOVE [19], WORD2VEC [17] et leurs variantes.¹⁶ Depuis la fin des années 2010, de nouvelles méthodes dites contextuelles et basées sur l’apprentissage profond, comme ELMo [20] et BERT [9], ont encore fait progresser ces techniques en améliorant les résultats sur de nombreuses tâches de traitement du langage naturel. Leur avantage principal est de produire un vecteur différent selon le contexte d’apparition du mot, et ce faisant de permettre une désambiguïsation à la volée, alors que les méthodes précédentes produisaient un seul vecteur par mot quel que soit son contexte.

16. Ces approches peuvent parfois se ramener à des techniques par factorisation de matrices [14].

A.2 Représentations vectorielles choisies

Pour notre analyse, nous considérons différentes représentations vectorielles que nous pensons représentatives des méthodes existantes :

- **tfidf** : vecteur des coefficients TF-IDF des mots du texte.
- **mangoes** : vecteur de mot produit par factorisation SVD de la matrice des coefficients PPMI.¹⁷
- **fasttext** [6] : vecteur de mot produit par FASTTEXT, une méthode neuronale étendant WORD2VEC qui permet de gérer les mots hors vocabulaire.¹⁸
- **bert** [9] : vecteur contextuel moyenne des quatre dernières couches d'un réseau BERT entraîné sur le français¹⁹. Pour une réponse comme pour une catégorie, nous avons moyenné les représentations des phrases la composant.

Les mêmes pré-traitements linguistiques ont été appliqués aux textes des réponses et des catégories, à savoir : une tokenisation, une normalisation de la casse, et la suppression des mots vides ("stopwords"). La représentation **tfidf** donne directement la représentation vectorielle d'un texte. Pour les autres méthodes, nous représentons un texte par la moyenne des vecteurs représentant les mots du texte. Outre sa simplicité, cette approche a l'avantage de très bien fonctionner en pratique [1]. Pour **bert**, chaque vecteur de mot est obtenu en moyennant les 4 dernières couches du réseau. Nous notons que les représentations **mangoes** et **fasttext** ont été entraînées sur un dump textuel du wikipedia français, tandis que **bert** a été entraîné sur un large corpus de pages web en français. Nous avons considéré d'autres variantes dans notre étude, sans que celles-ci changent les conclusions. Nous nous limitons donc à ce choix représentatif de représentations, toutes disponibles sur le site du projet.

En cohérence avec l'état de l'art, nous utilisons la distance cosinus, dont les valeurs sont comprises entre 0 et 1, pour mesurer la distance entre les représentations vectorielles des textes. Celle-ci a notamment l'avantage d'être moins sensible à des effets liés à la longueur des textes que la distance euclidienne. Des distances statistiques, telles que la distance de Wasserstein, sont également parfois utilisées [12]. Nous avons réalisé des expériences avec d'autres distances sans que ceci ne change les conclusions de notre étude.

A.3 Limites de l'approche

Notre approche, bien que s'appuyant sur des méthodes bien éprouvées de TAL, repose sur d'importantes hypothèses simplificatrices qui sont d'ailleurs très probablement partagées par l'approche mise en place par QWAM. Tout d'abord, les représentations vectorielles ont été calculées par entraînement sur des corpus généralistes et une piste d'amélioration serait d'affiner les représentations avec des corpus plus proches du domaine considéré. À l'exception de **bert**, les représentations choisies ignorent l'ordre des mots et les informations syntaxiques car les textes sont vus comme des "sacs de mots". Une piste d'amélioration serait

17. Représentations calculables avec <https://gitlab.inria.fr/magnet/mangoes>

18. Représentations, basées sur [10], reprises de <https://fasttext.cc/docs/en/crawl-vectors.html>

19. Représentations calculées avec le réseau camembert-base : <https://camembert-model.fr/>

de prendre en compte certains éléments syntaxiques et sémantiques comme, par exemple, le traitement des négations. Une autre piste serait de considérer les spécificités des énoncés de type questions–réponses, notamment l’existence de différents types de questions (questions oui-non, *wh*-questions), le fait que de nombreuses réponses sont des ellipses ou des *fragments* (à savoir, des énoncés dont la forme syntaxique est dégénérée et dont l’interprétation ne peut se faire indépendamment de la question). À ce titre, on notera que dans l’approche proposée l’assignation d’une réponse R à une catégorie C se fait ici indépendamment de la question Q . Ce choix se justifie par le fait que la question est fixe pour chaque ensemble de paires (Q, R) à classer mais revient à ignorer le contenu lexical de la question. Toutes ces pistes d’améliorations sont des problèmes actuels de la recherche en TAL.

En plus de l’approche présentée dans l’article, nous avons également testé des approches qui permettent de découvrir automatiquement les catégories, notamment LDA et sa version hiérarchique. Là encore, il nous a été impossible, quels que soient nos choix de paramétrage et de nombre de catégories (et de sous-catégories) de retrouver des effectifs de répartition proches de l’analyse officielle. De plus, nous n’avons généralement pas été capables d’affecter une sémantique (avec un intitulé textuel) aux groupes trouvés par l’algorithme à cause d’une très grande diversité des réponses (et des mots clés) affectées aux catégories trouvées. Ceci souligne le problème posé par l’absence d’évaluation de la catégorisation proposée dans l’analyse officielle et renforce les doutes sur la cohérence des groupes de réponse.

B Comparaison des répartitions avec seuil unique

Dans cet appendice, nous poursuivons l’analyse basée sur un seuil unique décrite en Section 3.2.1 en comparant les taux de répartition entre l’analyse officielle et notre méthode sur l’ensemble des questions. Une étude statistique des différences confirme que les pourcentages de répartition obtenus varient de manière significative selon le choix de la représentation, et ne correspondent pas à ceux de la synthèse officielle. Ceci est résumé par la Table 3, qui donne les écarts entre les répartitions obtenues par nos différentes approches et celles de la synthèse officielle. L’écart le plus faible est avec les représentations **tfidf** (l’approche la plus simple de notre étude), mais ces écarts sont élevés en moyenne pour toutes les représentations utilisées. La différence maximale est obtenue sur la même question pour toutes les méthodes : il s’agit de la question 32 du thème DC « Que proposez-vous afin de répondre à ce défi qui va durer ? » (à propos du “défi migratoire”). Le fait que cette question a un très fort taux de réponses vides et 10 catégories qui ont toutes un faible effectif explique ces différences très importantes. La différence minimale est obtenue, pour toutes les méthodes, sur la question 16.6 du thème ORG « Si vous avez été amené à demander un remboursement de soins de santé, pouvez-vous indiquer les éléments de satisfaction et/ou les difficultés rencontrés en précisant, pour chaque point, l’administration concernée ». Cette question, bien que comportant également un fort taux de réponses vides, ne comporte que deux catégories à la sémantique bien distinctes (satisfaction et insatisfaction) et aux effectifs équilibrés.

	tfidf	mangoes	fasttext	bert
Valeur moyenne	0.757	1.330	1.839	1.953
Écart type	0.676	1.532	2.253	2.118
Valeur minimale	0.270	0.195	0.096	0.155
Valeur maximale	4.713	12.389	18.952	17.853

TABLE 3 – Statistiques des distances L1 normalisées entre les répartitions par catégories de la synthèse officielle et les répartitions obtenues avec chacune de nos approches. La distance L1 normalisée pour une question donnée est obtenue en sommant la valeur absolue des écarts des répartitions pour chaque catégorie de la question, puis en divisant ce résultat par la somme des répartitions de la synthèse officielle pour cette question afin de tenir compte du nombre de réponses affectées aux catégories.

C Comparaison des répartitions avec seuils différents

Dans cet appendice, nous détaillons l’analyse basée sur des seuils par catégories décrite brièvement en Section 3.2.2. Pour mesurer l’agrément entre deux groupes de réponses U et U' affectées à une même catégorie C , nous définissons le taux de recouvrement

$$\frac{|U \cap U'|}{|C|}$$

où $|U \cap U'|$ est le nombre de réponses communes aux deux groupes et $|C|$ est le nombre de réponses attendues pour la catégorie C . Cette quantité est nulle lorsque les deux groupes sont disjoints, et vaut 1 lorsque les deux groupes sont identiques (les deux méthodes associent les mêmes réponses à cette catégorie).

La Figure 2 représente, pour quatre exemples de questions, les taux de recouvrement entre groupes de réponses affectées aux catégories pour notre méthode avec différentes représentations. Nous observons sans surprise que ces différentes approches sont généralement d’accord sur les réponses à associer aux catégories particulièrement simples : c’est ici le cas des catégories comme ‘aucun’ ou ‘tous’ dans le cas des questions de la colonne de droite, pour lesquelles les taux de recouvrement sont proches de 1. A l’inverse, pour les catégories ‘autres contributions’ qui servent de chapeau à des sous-catégories non liées entre elles, notre méthode d’affectation est sans doute mal adaptée et les 3 méthodes classent des réponses différentes dans ces catégories (les scores de recouvrement sont souvent proches de 0). En dehors de ces cas particuliers, nous observons une tendance nette (confirmée sur l’ensemble des questions) liée à la taille des catégories. En effet, le taux de recouvrement entre méthodes est relativement bon pour les catégories avec beaucoup de contributions et se dégrade avec la taille de la catégorie (sauf si la catégorie est très simple, voir point précédent). La cause de cet effet est difficile à identifier car il peut être dû à plusieurs facteurs : on peut penser à la difficulté intrinsèque à identifier un faible nombre de contributions parmi un grand nombre, mais aussi au fait que les catégories à faible effectifs puissent correspondre à des questions plus pointues et donc difficiles à classer, ou encore que le choix des catégories à faible

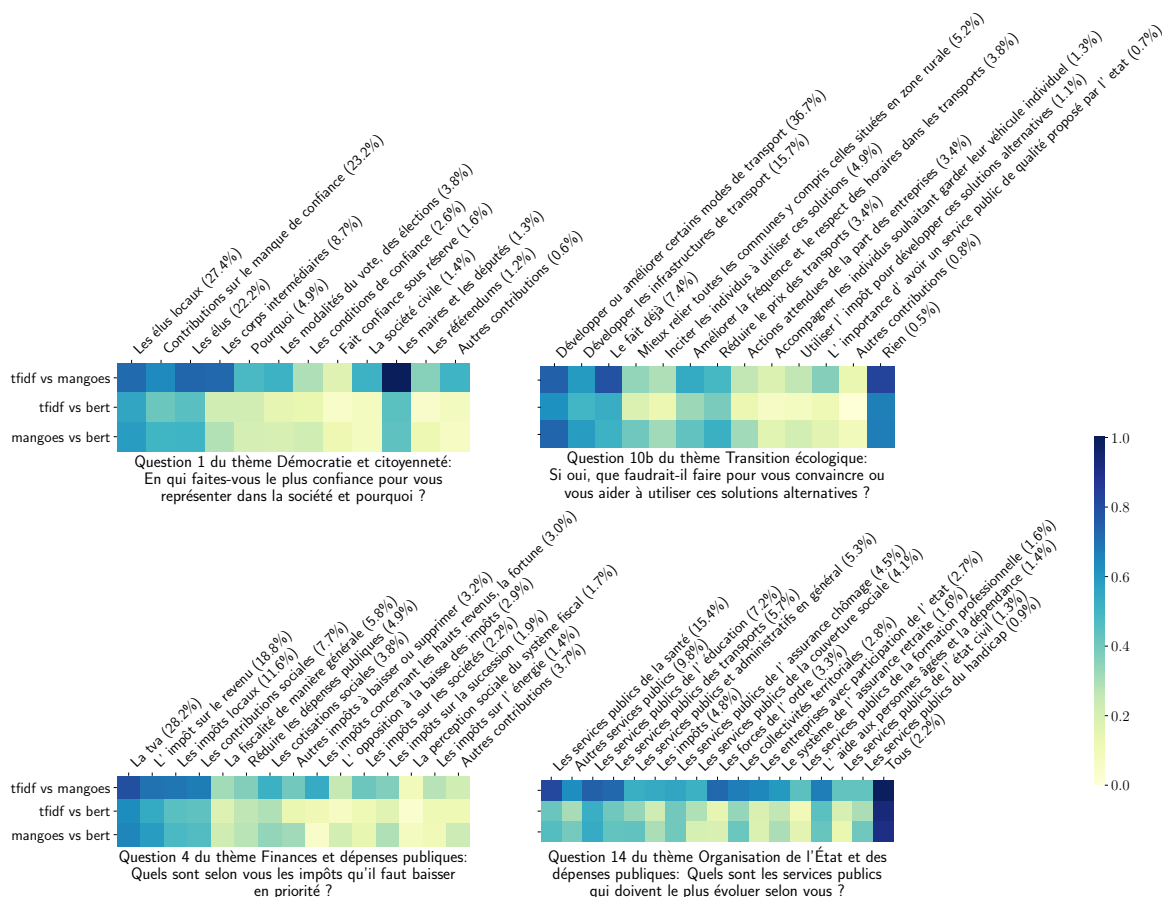


FIGURE 2 – Taux de recouvrement entre les réponses affectées à chaque catégorie par les méthodes **tfidf**, **mangoes** and **bert** pour quatre questions (une de chaque thème).

effectif dans la synthèse officielle ainsi que leurs effectifs sont potentiellement biaisés (nous reviendrons sur ce point plus loin).

Ces taux de recouvrement nous permettent d'identifier et d'explorer des catégories pour lesquelles la méthode de vectorisation des textes influence la classification des réponses. On observe ainsi des différences intéressantes dans les affectations produites par les différentes représentations vectorielles, en particulier entre **tfidf** d'une part et **mangoes** et **bert** d'autre part. Celles-ci s'expliquent par le fait que les premières requièrent que certains termes soient explicitement partagés entre la contribution et la catégorie (étendue à ses sous-catégories), alors que les secondes capturent des proximités sémantiques entre des mots différents. À titre d'exemple, pour la première question du thème TE, les représentations distribuées fournies par **mangoes** et **bert** permettent ainsi d'affecter la réponse « les avions » à la catégorie « la pollution », dont une des sous-catégories est « la pollution émanant des transports », en vertu de la proximité, correctement capturée par ces méthodes, entre les

termes *avions* et *transports*.