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

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)

29 mai 2020

Résumé

Face aux limites de la démocratie représentative, les consultations numériques participatives publiques permettent de solliciter, à différents niveaux de pouvoir, des contributions ouvertes de citoyens pour essayer de mieux impliquer les individus dans les décisions politiques. Si leur conception et leur mise en place posent des problèmes relativement bien connus, nous nous intéressons dans cet article aux enjeux liés à l'analyse automatique de contributions en langage naturel par des méthodes issues de l'intelligence artificielle. Il s'agit là d'un problème difficile pour lequel il existe des méthodes nombreuses et variées. En considérant comme cas d'étude les contributions aux questions ouvertes du Grand Débat National et l'analyse officielle qui en a été produite, nous montrons qu'il est impossible de reproduire les résultats de cette analyse et que différentes méthodes mènent à des résultats différents. Nous identifions également des choix arbitraires non explicités qui conduisent à émettre des doutes sur certains résultats de l'analyse officielle. Notre étude [1] met en lumière la nécessité d'une plus grande transparence dans l'analyse automatique de consultations ouvertes pour assurer la confiance du public dans leur restitution. Nous concluons par des pistes d'amélioration pour les consultations futures, afin qu'elles puissent être des outils utiles au débat public.

1 Introduction

La question des modes de représentation des citoyens possède une longue histoire en politique [12]. 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 représentations participatifs permettant d'impliquer plus directement et plus rapidement les individus dans les prises de décisions politiques [4, 2]. Ces dernières années ont ainsi vu l'émergence de plates-formes numériques participatives, qui servent en outre

1. Les données et le code source sont disponibles sur `https://gitlab.inria.fr/magnet/gdn-lib` pour permettre la réplication des expériences.

des objectifs d'e-gouvernance [2] [19]. 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'économie [3]. 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 pratique de ce type de consultations pose cependant des défis importants, à commencer par le fait d'assurer une bonne représentativité des contributeurs ainsi que des modes de questions et d'expression à même de refléter la pluralité des opinions. Au-delà de ces premiers enjeux qui concernent la conception et le mode de déploiement de la consultation, l'utilisation et le choix de certaines technologies pour le traitement, l'analyse et la restitution des réponses recueillies débouchent sur un second type d'enjeux, liés cette fois à des notions de réplicabilité et de transparence. L'apport de solutions à ces enjeux, sur lesquels se concentre cet article, est une condition essentielle à l'avenir de telles consultations.

Les consultations numériques participatives sont, en général, réalisées par Internet et s'appuient parfois sur des réseaux sociaux. Elles nécessitent donc l'utilisation de technologies de l'information pour la collecte et le stockage mais aussi pour l'aggrégation statistique et la visualisation des données issues de ces plateformes, qui sont typiquement de grande taille et ne peuvent être traitées exhaustivement par des humains. Ces consultations étaient jusque récemment limitées à des questionnaires fermés, c'est-à-dire avec des choix de réponses encadrés. Mais des techniques récentes issues de l'Intelligence Articifielle (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 et en grande quantité par les citoyens [4]. 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 ne va pas sans risques. En effet, l'analyse automatique de textes libres est une tâche complexe qui ne se limite pas à un simple comptage de réponses, et dont les résultats sont difficiles à évaluer de manière objective. Afin d'assurer une forme de confiance du public dans la restitution des consultations, il est donc 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) [5], la consultation citoyenne initiée en janvier 2019 par le président de la République française.

2. "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." (`http://portal.unesco.org/ci/fr/ev.php-URL_ID=3038&URL_DO=DO_TOPIC&URL_SECTION=201.html`)

3. 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`

4. Voir [6] pour un exemple de déploiement de ce type de méthodes sur des consultations issues de la plate-forme québécoise Votepour.ca.

5. `https://granddebat.fr/`

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 suscitée plusieurs dizaines de milliers de réponses textuelles parfois longues et argumentées. Le traitement automatique de ces contributions textuelles est l'objet principal de notre étude.

L'analyse de cette consultation en ligne a été confiée à OpinionWay [7] qui a sous-traité l'étude des questions ouvertes à l'entreprise Qwam [8]. Les auteurs de l'analyse officielle ont choisi, pour chaque question ouverte, de déterminer des catégories et de fournir des pourcentages de répartition des réponses dans ces catégories. Pour réaliser cette analyse, Qwam a utilisé des méthodes automatiques internes et un post-traitement humain. Les résultats fournis prennent la forme de taux de répartition dans des catégories et sous-catégories, elles-mêmes définies par des intitulés textuels courts comme illustré dans l'exemple de la Figure 1.

Malheureusement, l'ensemble de la chaîne de traitement mise en place par Qwam et OpinionWay pour analyser les réponses aux questions ouvertes manque cruellement de transparence, malgré les demandes [10] et les annonces [9] 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. Ces méthodes peuvent mener à des choix de catégories différents et à des répartitions différentes, et ces différences peuvent être accentuées par les traitements humains réalisés au cours de l'analyse. Les affectations individuelles des réponses aux catégories n'ont pas non plus été rendues publiques. Cette opacité sur les méthodes utilisées, le code et les résultats empêche toute validation externe, et a fortiori toute réplication.

À défaut de pouvoir répliquer la méthode d'analyse qui a mené à la synthèse officielle, et de pouvoir la valider sur des assignations précises des réponses aux catégories fournies, nous proposons dans cet article une rétro-analyse qui met en lumière le besoin de transparence et vient nuancer les conclusions de l'analyse officielle. Plus spécifiquement, en nous appuyant sur différentes méthodes de l'état de l'art en TAL, décrites ici exhaustivement, nous proposons différentes approches pour (ré-)affecter les textes des contributions aux catégories et sous-catégories proposées par 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. Nos résultats mettent en évidence des questions difficiles à analyser, souvent avec des réponses longues et argumentées. Une étude qualitative de ces questions nous permet

6. https://www.nouvelobs.com/politique/20181219.OBS7366/grand-debat-national-voulu-par-macron-a-quoi-faut-il-s-attendre.html

7. www.opinion-way.com

8. http://www.qwamci.com/

9. Voir le tweet du secrétaire d'État chargé du numérique : https://twitter.com/cedric_o/status/1115171491384037377

en outre d'identifier des biais dans le choix des catégories de la synthèse officielle, d'émettre des doutes sur les taux présentés, et de formuler des idées d'analyses complémentaires.

Notre étude nous amène à formuler des suggestions de bonnes pratiques en vue de consultations futures. Nous insistons évidemment sur la nécessité de transparence, notamment par la publication du code et des paramétrages, ou à défaut des affectations individuelles des contributions, afin de permettre une validation externe par des experts et des citoyens et de favoriser ainsi la confiance du public dans les restitutions. 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. Nous discutons enfin de la manière dont une supervision humaine partielle peut efficacement guider l'analyse automatique.

Le reste de l'article s'organise comme suit. La Section 2 présente les données du Grand Débat National et la synthèse officielle, puis discute des limites de cette dernière. La Section 3 présente notre méthodologie et les résultats de notre rétro-analyse. Nous concluons par une discussion sur les bonnes pratiques 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 y ont été fournies par les contributeurs et la synthèse qui en a été faite 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 préfini 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

	Questions	Contributions			Catégories de la synthèse	
Thème	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.

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 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.

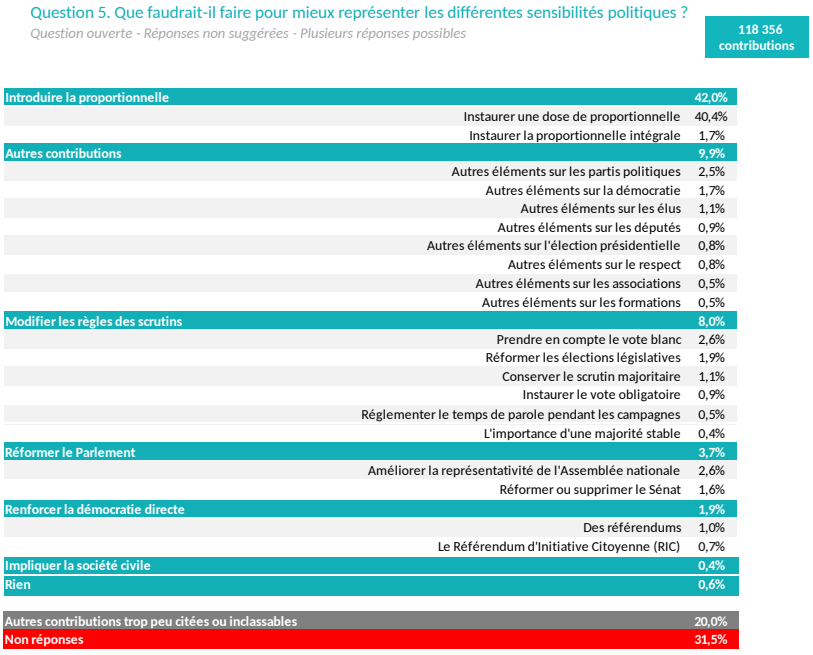
En ce qui concerne la longueur des questions et des contributions, mesurées en nombre de mots, notons tout d'abord que les questions ont une taille moyenne de 17.3 mots tout thème confondu. La taille des contributions, quant à elle, varie aussi très fortement selon les questions. Ainsi, la longueur moyenne 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 mesures de base montrent que les participants ont joué le jeu de la consultation ouverte et ont exprimé des opinions détaillées dans leurs réponses. Ceci souligne l'importance qu'aurait pu avoir un retour vers les participants pour montrer comment avaient été traitées leurs réponses individuelles.

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 gouvernement [10]. La synthèse officielle [11], produite en avril 2019, contient deux volets, correspondant à des sous-ensembles de questions distincts et à des prestataires différents. Le premier volet consiste

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`



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 ».

en un ensemble de statistiques descriptives compilées par OPINIONWAY à partir des réponses aux 52 questions fermées. Ces enquêtes d'opinion et les méthodes d'analyse statistique associées sont largement répandues et l'on en connaît assez bien les avantages et les inconvénients : les biais potentiels dus aux choix des questions et, pour les enquêtes aux contributions anonymes en ligne, le problème des répondants multiples et de la représentativité des répondants. Nous ignorons ce premier volet pour nous intéresser spécifiquement au second volet qui traite des questions ouvertes.

Celui-ci a été réalisé par l'entreprise QWAM 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 [12], de doublons de participants mais aussi de doublons de réponses (parfois suggérées par des associations), et de représentativité. Nous laissons de côté ces problèmes pour nous concentrer sur l'analyse des 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,

12. D'ailleurs relevés par certains participants avec des réponses comme « *C'est de la foutaise, toutes les questions sont orientées !!! On est pas là pour répondre à un QCM !"* », voir `https://granddebat.fr/projects/la-transition-ecologique/collect/participez-a-la-recherche-collective-de-solutions-1/proposals/cest-de-la-foutaise-toutes-les-questions-sont-orientees-on-est-pas-la-pour-repondre-a-un-qcm-1`

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 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 », ..., « Instaurer le vote obligatoire ».

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 (voir par exemple la catégorie « Autres contributions » ou encore la sous-catégorie « L'importance d'une majorité stable »). Pour l'ensemble des questions, il existe aussi deux catégories particulières : les « non-réponses », la catégorie qui comptabilise les textes vides, et les « inclassables », qui sont les réponses n'ayant pu être rangées dans aucune des catégories choisies par l'analyse officielle. La longueur moyenne des intitulés de catégories est de 5.5 mots, alors que ceux 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. Les taux reportés dans l'étude nous permettent de déduire qu'une réponse d'une catégorie peut être classée dans une seule sous-catégorie, plusieurs sous-catégories, voire aucune sous-catégorie. En effet, 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 également 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 $42 + 9,9 + 8 + 3,7 + 1,9 + 0,4 + 0,6 = 66,5\%$. Ce recouvrement entre catégories peut sembler relativement faible, alors que les intitulés des catégories et sous-catégories suggèrent la possibilité d'affectations multiples des contributions.

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.

Avant de nous intéresser à la méthode employée pour obtenir cette synthèse, on peut d'ores et déjà émettre certains constats critiques. Tout d'abord, il est important de remarquer

que, même si l'objectif de catégorisation semble pertinent, il aurait pu être complété par d'autres analyses comme l'analyse des sentiments ou la détection d'idées émergentes. Ensuite, comme nous l'avons déjà signalé, les choix des (intitulés des) catégories et sous-catégories et le faible recouvrement des répartitions pour certaines questions interrogent. Mais nous souhaitons particulièrement noter 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, bien que, comme nous venons de le souligner, les participants aient fait l'effort de rédiger des réponses parfois longues aux questions.

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 machine, 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. 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. Plus généralement, le problème est rendu difficile par la nature textuelle, donc non structurée, des données. Ces difficultés sont exacerbées par 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 de non-coopérativité (au sens des maximes de Grice), 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 machine, 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

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

clustering spectral, leurs extensions hiérarchiques, ainsi que les algorithmes de classification supervisée (comme Naive Bayes, 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 [13] pour un survol). Plus spécifiquement orientée vers le clustering de données textuelles, Latent Dirichlet Allocation (LDA) [3] existe en version plate ou hiérarchique et contient une réduction de dimension implicite à travers la modélisation par variable latente 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 de manière plus détaillée sur ce point dans la Section 3.1.2).

Parmi ce large choix de techniques et de paramétrages associés, quelle est donc la méthode retenue par l'entreprise Qwam ? Selon le rapport d'étude [14], 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 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 et sous-catégories et des intitulés textuels associés n'est pas clairement 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 par des humains (par exemple, sous la forme d'un score d'accord entre humains), n'est fournie pour garantir la validité des affectations réponses-catégories.

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 Qwam et OpinionWay est parfaitement opaque. Nous ne disposons même pas d'élément permettant de retrouver à quelle catégorie est affectée une

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

réponse particulière dans l'analyse officielle. Il est donc impossible de vérifier si une autre approche fournit des résultats identiques ou similaires : nous ne pouvons travailler que sur les 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 répliquer 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 notre approche générale ainsi que les différentes représentations vectorielles de texte utilisées. Les données, les représentations vectorielles et le code source pour répliquer 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. 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 générale

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.

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. Au-delà de la sélection du seuil, que nous discutons dans nos expériences, le choix essentiel est celui de la distance entre textes, qui dépend à son tour de la représentation choisie pour encoder les contenus textuels. En cohérence avec l'état de l'art, nous allons nous appuyer sur une distance cosinus entre des représentations vectorielles des textes (c-à-d, une mesure de l'angle entre les vecteurs représentant 2 textes). Celle-ci a notamment l'avantage d'être moins sensible à des effets liés à la longueur des textes que la distance euclidienne. [15] Les représentations vectorielles que nous considérons sont présentées dans la section suivante.

3.1.3 Représentations vectorielles du texte

Les représentations vectorielles des mots et des textes possèdent une longue histoire en analyse automatique, remontant au moins aux travaux fondateurs de Salton [18] en recherche d'information. Les modèles vectoriels ont aussi largement imprégné les recherches en TAL, d'abord dans le cadre de la sémantique distributionnelle [20] puis plus récemment dans le contexte des *word embeddings* (ou *plongements lexicaux*) neuronaux [15].

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

15. Des distances statistiques, telles que la distance de Wasserstein, sont également parfois utilisées [9].

tout d'abord sur des méthodes classiques de réduction de dimensions par factorisation de matrices, notamment l'analyse en composantes principales, puis sur des méthodes par apprentissage neuronal depuis les années 2010, telles que GloVe [16], Word2vec [14] et leurs variantes. [16] Depuis la fin des années 2010, de nouvelles méthodes dites contextuelles et basées sur l'apprentissage profond, comme ELMo [17] et Bert [7], 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.

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** [5] : vecteur de mot produit par Fasttext, une méthode neuronale étendant Word2vec qui permet de gérer les mots hors vocabulaire. [18]
- **bert** [7] : vecteur contextuel extrait d'un réseau de neurones profond. [19]

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 de "stop words". 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.

Enfin, comme évoqué plus haut, la distance entre deux représentations vectorielles de textes est mesurée par la distance cosinus (dont les valeurs sont comprises entre 0 et 1), qui fait office de référence dans ce domaine. Ici encore, nous avons réalisé des expériences avec d'autres distances sans que ceci ne change les conclusions de notre étude.

3.1.4 Limites de l'approche

Avant de discuter les résultats, notons que notre approche, bien que s'appuyant sur des méthodes bien éprouvées de TAL et de recherche d'information, repose sur d'importantes

16. Ces approches peuvent parfois se ramener à des techniques par factorisation de matrices [11].
17. Représentations calculables avec `https://gitlab.inria.fr/magnet/mangoes`
18. Représentations, basées sur [8], reprises de `https://fasttext.cc/docs/en/crawl-vectors.html`
19. Représentations calculées avec le réseau camembert-base : `https://camembert-model.fr/`

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%
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 ? ».

hypothèses simplificatrices — certaines de celles-ci sont d'ailleurs très probablement partagées par l'approche mise en place par OPINIONWAY et QWAM. Tout d'abord notons que les représentations vectorielles ont été calculées par entraînement sur des corpus généralistes et qu'une piste d'amélioration serait d'affiner les représentations avec des corpus de domaines. On notera également que les représentations choisies ignorent l'ordre des mots et les informations syntaxiques (les énoncés R et C sont vus comme des "sacs de mots"). Ici encore, une piste d'amélioration serait de prendre en compte explicitement certains éléments syntaxiques et sémantiques comme le traitement des négations mais aussi 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.

Nous notons que la plupart des pistes d'améliorations évoquées ci-dessus restent des problèmes largement ouverts en recherche en TAL.

3.2 Résultats

3.2.1 Seuil unique basé sur le taux d'inclassables

Comme expliqué en Section 3.1.2, une fois la représentation vectorielle choisie, l'affectation d'une réponse R à une catégorie C se fait sur le base d'un seuillage de la distance cosinus entre R et C. Dans cette première expérience, nous choisissons *un seuil unique* de

façon à obtenir le même taux d'inclassables que dans l'analyse officielle. Nous classons alors une réponse dans une catégorie donnée si sa distance à cette catégorie est inférieure au seuil. Nous obtenons ainsi, pour chacune des représentations vectorielles, des taux de répartition des réponses dans les catégories choisies par 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 fixant le taux d'inclassables à 13.2%, nous obtenons les répartitions présentées dans la Table 2. Un premier constat est que notre méthode a tendance à sur-évaluer les taux de répartition par rapport à la synthèse officielle, avec un grand nombre de contributions affectées à plusieurs catégories. À la vue des intitulés des catégories, ces affectations multiples sont plausibles et il est probable qu'elles soient trop limitées dans la synthèse officielle. Par exemple, la catégorie « Les modalités du vote, des élections » est certainement sous-estimée. Avec notre méthode, elle est représentée par le texte « Les modalités du vote, des élections, le suffrage universel, la proportionnelle, la confiance en la représentativité, une représentativité plus forte garantirait la confiance » (intitulé de la catégorie plus ceux des sous-catégories) et nous obtenons des taux qui sont sans doute plus proches de la réalité.

Une deuxième observation, à partir de la Table 2, est que les différences de répartition sont importantes entre l'analyse officielle et notre méthode, ce qui se vérifie pour l'ensemble des questions. En effet, 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 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.

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* de façon à forcer une répartition identique à celle de l'analyse officielle. Les résultats obtenus avec différentes méthodes 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,

	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.

il nous est impossible de considérer l'analyse officielle car les affectations individuelles des réponses aux catégories ne sont pas connues. Nous allons donc mener cette étude en considérant notre méthode d'affectation dans les catégories avec nos différentes représentations vectorielles.

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

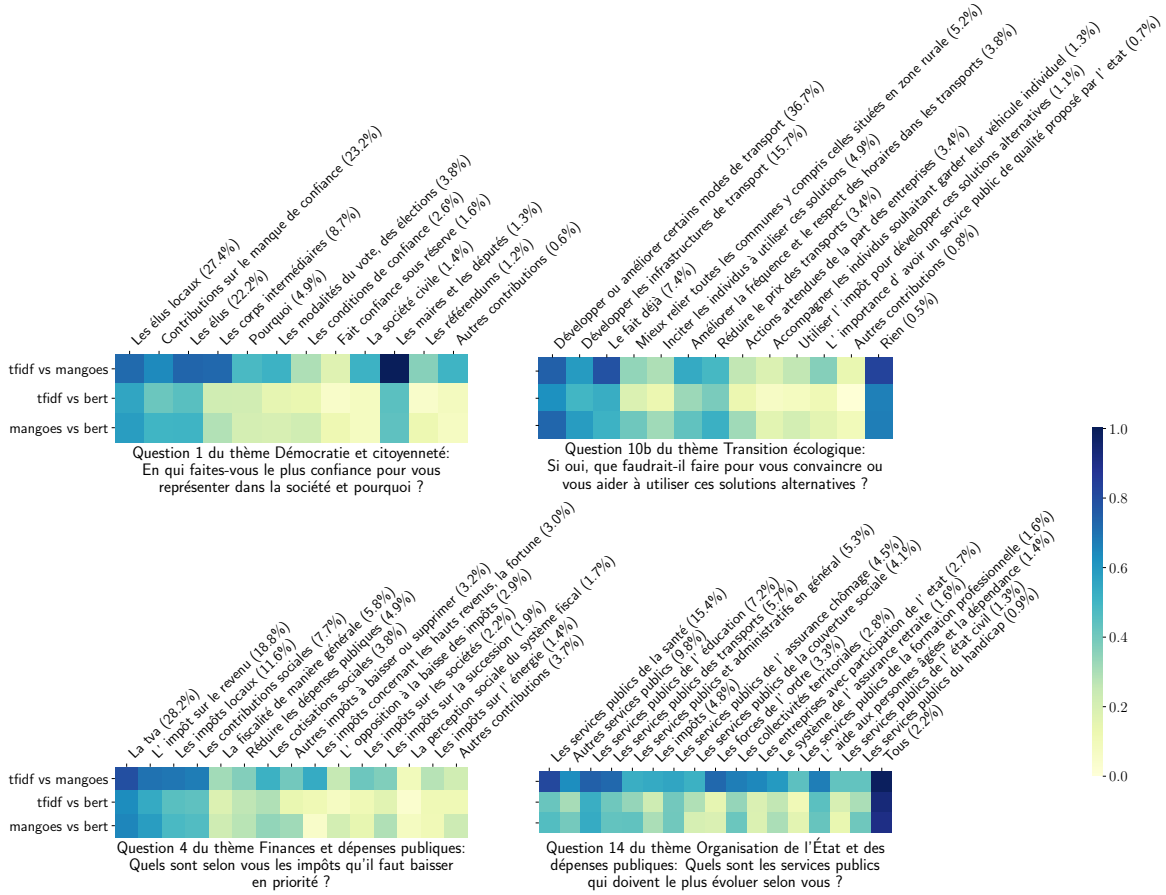


FIGURE 2 – Taux de recouvrement entre les réponses affectées à chaque catégorie par les méthodes **tfidf**, **mangoes** and **bert** pour quatre questions (une de chaque thème).

plus pointues et donc difficiles à classifier, ou encore que le choix des catégories à faible 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 distributionnelles 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*.

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. Rappelons que le seuil permet d'affecter une réponse dans la catégorie si la distance entre la réponse et la représentation textuelle de la catégorie est inférieure au seuil. Dans une troisième expérience, nous avons réalisé une étude statistique de la répartition des seuils. Celle-ci 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 nous a permis d'identifier certaines anomalies dans les répartitions de la synthèse officielle. Nous les détaillons ci-dessous.

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 quelques 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 ». Cette catégorie se décompose en 14 sous-catégories « Dont les acteurs locaux », « les communes », ..., « les élus locaux ». 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 malgré tout classé manuellement les réponses dans la catégorie acteurs publics et trouvé un taux de 54.5%. Ceci montre que la catégorie est largement sous-estimée par l'analyse officielle avec un taux de 43.4%, soit une différence de l'ordre de 15 000 réponses entre affectation manuelle et l'analyse officielle. Si on considère 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.

3.2.4 Anomalies dans le choix des catégories de la synthèse officielle

Notre 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 a également permis de relever des biais dans le choix des catégories. En effet, certaines réponses concernent la prise en charge par l'individu lui-même des solutions alternatives. Nous avons donc choisi de considérer une catégorie « prise en charge par l'individu », catégorie qui n'apparaît pas dans les catégories de l'analyse officielle. En annotant manuellement les réponses, nous trouvons 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 », « les citoyens », « je suis grand », ..., « c'est mon problème », ou plus longues comme « les francais sont assez intelligents pour les trouver seuls », ..., « les citoyens sont les premiers maitres 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 montre l'aspect arbitraire du choix des catégories, qui ignore ici des réponses ayant une sémantique forte pour l'étude : la nécessaire implication des individus dans les changements de comportement.

4 Conclusions et Perspectives

Dans cet article, nous avons présenté une rétro-analyse de la synthèse officielle des questions ouvertes du Grand Débat National. Notre étude s'appuie sur des méthodes de l'état de l'art en TAL pour ré-affecter les contributions 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 malgré son manque de transparence**, 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 d'OpinionWay et Qwam, nos résultats confirment 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 futures afin d'assurer une plus grande transparence et une meilleure fiabilité. En particulier :

- Il est nécessaire d'**introduire davantage de transparence dans le 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) et un débat contradictoire. Dans le cas du GDN, cela aurait pu prendre la forme d'une publication des affectations de chaque contribution individuelle aux catégories. Ceci, à défaut d'une prise en compte politique, permettrait, à minima, à chaque participant de savoir comment sa contribution a été prise en compte.
- Il est utile de **confronter différentes analyses**, dans les conditions de transparence indiquées ci-dessus. En effet, chaque méthode d'analyse automatique repose sur des hypothèses spécifiques, avec ses propres biais, et donne donc une interprétation possible du corpus. La confrontation de plusieurs analyses est utile pour nuancer certaines conclusions et peut mener à une synthèse finale plus fiable. Une répartition des contributions dans des catégories thématiques peut également être complétée par d'autres axes d'analyse, comme la recherche d'idées émergentes ou l'analyse de sentiments.
- Pour dépasser les limites intrinsèques des méthodes non supervisées, il est possible d'**introduire une supervision humaine partielle** qui permet de déployer des méthodes supervisées ou semi-supervisées plus fiables et dont les performances peuvent être évaluées de manière quantitative. Une piste prometteuse est de **concevoir des consultations plus collaboratives et interactives** : on peut penser notamment à une annotation partielle des textes par des volontaires (voir l'initiative de la Grande Annotation [20]), à une consultation en deux phases où les contributeurs ayant la possibilité de voir comment leurs contributions ont été prises en compte sont invités à signaler des problèmes permettant alors un nouvel apprentissage tenant compte des

20. `https://grandeannotation.fr/`

corrections, ou encore à un système permettant de voter pour les contributions ou de les commenter.

Ces pistes offrent des perspectives pour rendre les consultations participatives plus utiles au débat public. Elles permettent en particulier de donner des gages aux citoyens quant à la prise en compte de leur opinion, favorisant ainsi leur implication, qui est la condition indispensable à l'avenir de telles consultations.

Références

[1] Sanjeev Arora, Yingyu Liang et Tengyu Ma : A simple but tough-to-beat baseline for sentence embeddings. *In Proceedings of the International Conference on Learning Representations*, 2017.

[2] Marie-Hélène Bacqué et Yves Sintomer : *La démocratie participative : Histoire et généalogie*. La Découverte, 2011.

[3] David M Blei, Andrew Y Ng et Michael I Jordan : Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[4] Loïc Blondiaux : *Le Nouvel Esprit de la démocratie. Actualité de la démocratie participative*. Seuil, 2008.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin et Tomas Mikolov : Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[6] 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.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee et Kristina Toutanova : BERT : pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[8] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin et Tomas Mikolov : Learning word vectors for 157 languages. *In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[9] Matt Kusner, Yu Sun, Nicholas Kolkin et Kilian Weinberger : From word embeddings to document distances. *In International conference on machine learning*, pages 957–966, 2015.

[10] 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.

[11] Omer Levy et Yoav Goldberg : Neural word embedding as implicit matrix factorization. *In Advances in Neural Information Processing Systems 27 : Annual Conference on*

Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2177–2185, 2014.

[12] Bernard Manin : *Principes du gouvernement représentatif*. Flammarion, 3ème édition, 2019.

[13] Christopher D. Manning, Prabhakar Raghavan et Hinrich Schütze : *Introduction to Information Retrieval*. Cambridge University press, 2008.

[14] Tomas Mikolov, Kai Chen, Greg Corrado et Jeffrey Dean : Efficient estimation of word representations in vector space. *In 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean : Distributed representations of words and phrases and their compositionality. *In Advances in neural information processing systems*, pages 3111–3119, 2013.

[16] Jeffrey Pennington, Richard Socher et Christopher Manning : Glove : Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[17] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee et Luke Zettlemoyer : Deep contextualized word representations. *In Proc. of NAACL*, 2018.

[18] Gerard Salton, Anita Wong et Chung-Shu Yang : A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[19] Anthony Trotta : *Advances in E-Governance : Theory and Application of Technological Initiatives*. Productivity Press, 2017.

[20] Peter D. Turney et Patrick Pantel : From frequency to meaning : Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.