Les multiples agendas médiatiques des Gilets jaunes sur YouTube [version longue et de travail]

Bilel Benbouzid, Hervé Guérin

Les multiples agendas médiatiques des Gilets jaunes sur YouTube.

Exploration d'un corpus de vidéos avec les topics models.

Bilel Benbouzid Hervé Guérin

Le mouvement des Gilets jaunes est apparu comme une forme paroxystique des reconfigurations de l'écosystème médiatique (Cardon and Granjon, 2010; Ferron, 2019; Granjon, 2014a, 2018), par au moins trois aspects : une autonomie revendiquée des Gilets jaunes vis-à-vis des médias traditionnels qui rompt avec les situations observées classiquement de militantisme de *prime time*, c'est-à-dire de coopérations/conflits entre activiste et journaliste dans l'accès à l'espace médiatique pour la représentation des groupes mobilisés ; un soutien rapide et spontané d'une série de médias autonomes et alternatifs natifs du web qui ont permis de servir des intérêts sociaux correspondant à ceux que les Gilets jaunes entendent défendre ; une capacité à s'emparer des réseaux sociaux et y construire un dispositif d'échange horizontal, pour non seulement coordonner les actions sur le terrain des manifestations, sensibiliser à des situations d'injustice et faire circuler des cadres d'interprétation des problèmes, mais aussi, et surtout, pour débattre de la couverture médiatique accordée au mouvement, les Gilets jaunes se réfugiant ainsi sur Facebook dans une bulle protectrice de filtre de l'information.

Dans ce contexte d'évolution des modalités de production de l'espace médiatique, comment analyser la *médiatisation* du mouvement des Gilets jaunes en tenant compte de cette nouvelle économie de la représentation médiatique (Granjon, 2014b) ? Pour répondre à cette question, nous avons mené une analyse quantitative de la couverture médiatique des Gilets jaunes sur la plateforme YouTube. L'intérêt pour YouTube réside dans le fait que les canaux médiatiques de la plupart des acteurs qui composent l'espace public s'étendent désormais vers la plateforme : presse écrite, radio, télévision, média alternatifs, partis politiques, associations militantes, citoyen ordinaire etc. : autant d'acteurs présents sur YouTube qui sont mis sur le même plan dans l'espace médiatique par la possibilité de créer une chaîne et publier du contenu. Dans ce contexte, comment se forme l'agenda médiatique (McCombs and Shaw, 1972) selon les types d'acteurs qui composent l'espace médiatique et politique sur YouTube ? Et quelles ont été les évolutions de cet agenda durant les huit premiers mois du mouvement ?

Si Facebook est sans conteste l'espace numérique de la contestation des Gilets jaunes, YouTube est celui d'où l'on peut observer le mieux la lutte médiatique entre une pluralité de médias pour imposer un sens légitime à donner à l'événement. C'est en effet un terrain d'enquête commode pour opérer une analyse quantitative de contenu : les discours produits peuvent être agrégés pour représenter l'écosystème médiatique dans son ensemble à partir d'un corpus homogène, dans le sens de contenu partageant un support de communication commun, en l'occurrence la vidéo (explorée comme une source textuelle à partir des sous-titres). C'est ici que réside l'intérêt de notre enquête vis-à-vis des quelques analyses quantitatives de contenus réalisées sur le vif et visant à analyser les couvertures médiatiques des Gilets jaunes (Sebbah et al., 2018; Sebbah, Loubère, et al., 2019, 2019;

Sebbah, Marchand, et al., 2019). Pour intégrer les reconfigurations de l'écosystème médiatique, les chercheurs ont tenté de rendre compte des tensions entre la parole citoyenne des Gilets jaunes sur les réseaux sociaux et l'information journalistique. Ces études ont donc comparé des contenus produits sur différents espaces, mais l'hétérogénéité des corpus collectés implique de dissocier les explorations selon les espaces médiatiques car les contenus restent incommensurables - des journaux télévisés et des tweets dans une enquête de l'INA (Poels and Lefort, 2019) et des articles de presse papier et des discussions sur Facebook dans une série de rapports du LERASS mentionnés plus haut (voir la synthèse dans (Souillard et al., 2020)). Cette dissociation rend difficile la comparaison des couvertures médiatiques. En revanche, YouTube offre la possibilité de comparer des contenus produits par des types d'acteurs différents qui se sont engagés sur la plateforme dans une lutte pour la représentation médiatique.

Une enquête quantitative sur YouTube n'est possible que si deux conditions sont respectées : avoir accès à un corpus clairement délimité, d'une part, et à des outils taillés sur mesure pour en assurer l'exploration, d'autre part. La première condition est satisfaite si on parvient à arrêter un corpus de vidéos à partir d'une liste de classes d'acteurs (des chaînes YouTube) suffisamment étendue pour représenter l'espace public élargi - des médias traditionnels jusqu'aux youtubeurs vulgarisateurs, en passant par la sphère contre-informationnelle et les Gilets jaunes eux-mêmes. La seconde se trouve du côté des outils de fouille textuelle. Nous avons choisi d'utiliser les *topics models*, une méthode développée dans le domaine de la linguistique computationnelle depuis plus d'une dizaine d'années, devenue classique dans le domaine de l'analyse quantitative de contenu en SHS (Cointet and Parasie, 2018; Evans and Aceves, 2016; Wesslen, 2018), qui permet, de manière non supervisée, de modéliser statistiquement quels ensembles de mots font un "topic" et dans quelle mesure une vidéo traite de ce topic. Après avoir présenté dans une premières partie la construction du corpus, puis dans une deuxième partie la méthode des *topics models*, nous présenterons dans un troisième temps les principaux résultats obtenus à partir d'une série d'analyses statistiques appliquées aux topics. Ces résultats nous permettrons de discuter des différents agendas médiatiques du mouvement des Gilet jaunes. En guise de conclusion, nous revenons sur les enjeux de l'analyse quantitative de contenu comme *statactivisme* (Bruno and Didier, 2014) dans les débats autour des biais de représentation médiatique des mouvements sociaux.

Corpus

Le lieu d'extraction de notre corpus est particulier et assez différent de ceux explorés habituellement dans les analyses des traitements médiatiques des mouvements sociaux. Il ne s'agit pas d'un dossier de presse, ni d'un ensemble de document d'archive INA de journaux télévisés ou d'un jeu de données tiré des plateformes Twitter ou Facebook, ces dernières étant devenues des terrains d'enquête classiques pour l'étude des mouvements sociaux.

Ce choix de YouTube repose sur une conception de l'espace public élargi qui met sur le même plan des formes de communication de nature très différentes. Il nous faut donc commencer par expliquer la manière dont nous avons conçu la sphère médiatique et

politique sur YouTube. Ensuite, sur un autre registre, ce corpus a une valeur patrimoniale pour les sciences humaines et sociales. C'est pourquoi notre indexation des chaînes, dont nous explicitons plus bas la taxonomie en sept catégories, est entièrement livrée en annexe. Cette livraison a ainsi pour objectif d'amorcer des études YouTube ultérieures reposant sur un espace de commune mesure permettant la confrontation des énoncés scientifiques sur les traitements médiatiques des mouvements sociaux. Enfin, la nature audiovisuelle du matériau collecté, traité comme du contenu textuel, pousse à l'explicitation des techniques de délimitation et de préparation du corpus. Les techniques employées, bien qu'algorithmiques, relèvent d'un savoir-faire artisanal, rarement consigné dans les manuels, que nous souhaitons tout simplement transmettre dans cet article.

La sphère médiatique et politique sur YouTube

Dans cette enquête, nous avons souhaité représenter l'espace médiatique et politique produit sur YouTube pour la France sans limiter notre corpus aux seules chaînes de production journalistique de l'information. Ce critère d'ouverture tient à la spécificité de l'espace numérique : dans un contexte où la distinction entre médias traditionnels et « médias sociaux » est de moins en moins évidente, comment les différents formats de « communication » sont-ils liés les uns aux autres et qui parvient le mieux à définir l'agenda médiatique ? Pour être en mesure de répondre à cette question, nous avons étendu notre corpus à tous les types de chaînes produisant des opinions, des analyses et des décryptages ; nous avons aussi inclus les chaînes explicitement liées à des entités économiques, politiques, syndicales et associatives ; enfin, nous avons intégré les organisations de services publics qui gèrent sur YouTube leurs relations publiques.

Pour arriver à une telle hétérogénéité, nous avons construit des listes de chaînes a priori selon les types d'acteurs qu'il nous semblait important de pouvoir situer dans l'espace de YouTube : les médias professionnels ; les chaînes de youtubeurs notoires tournés vers la politique ; les associations militantes ; les députés ; les chaînes de candidats aux élections européennes ; les chaînes de partis politiques ; les chaînes créées à l'occasion des Gilets jaunes ; les chaînes d'associations tournées vers des causes publiques ; les chaînes de grandes institutions publiques ou privées – nous sommes ainsi parvenu à dix types d'acteurs dont nous avons cherché la présence sur YouTube.

En effet, pour chacune de ces catégories d'acteurs, nous avons manuellement dressé des listes de chaînes. Une fois agrégées, ces chaînes nous ont alors permis d'étendre notre corpus en les utilisant comme des points d'entrée pour dresser un réseau de chaînes plus large. Nous avons profité de l'effet boule de neige du réseau de recommandation entre chaînes (nous reviendrons plus bas sur cette fonctionnalité de YouTube). Par exemple, les chaînes régionales de France Télévision sont toutes liées les unes aux autres ou encore les chaînes de vulgarisation scientifique ont tendance à former un réseau de chaînes « amies ». Nous avons répété plusieurs fois cette opération qui s'apparente à une méthode par chaînage. Nous avons opéré plusieurs vagues de collecte en ajoutant systématiquement les nouvelles chaînes. Au fil du crawling, l'apport en chaînes nouvelles correspondant au thème défini s'est progressivement raréfié, jusqu'à s'épuiser aux alentours de 800 chaînes.

Cette procédure de collecte est néanmoins biaisée envers les individus les plus coopératifs sur YouTube. De nombreuses chaînes n'indiquent aucune chaîne amie ni ne sont recommandées par d'autres. Pour limiter ce biais, nous avons cherché de nouvelles chaînes dans la base de données de Wizdéo en consultant les listes de chaînes similaires associées à chacune de nos chaînes – une métrique de similarité est calculée à partir des commentateurs communs. Cette procédure fastidieuse a néanmoins permis d'atteindre un effectif total de 1400 chaînes.

Mais de quoi cette liste de 1400 chaînes est-elle la représentation ? Il ne s'agit pas d'un échantillon représentatif car certaines catégories de chaînes apparaissent de manière quasi exhaustive, alors que d'autres sont sous-représentées. En effet, si pour certaines catégories il est apparu simple de constituer une liste quasi-exhaustive (chaînes des députés, des ministères, des médias traditionnels), la limite du corpus est beaucoup plus difficile à arrêter lorsqu'il s'agit de petites associations militantes ou de vlogueurs individuels publiant des vidéos pour faire valoir leur opinion sur l'actualité. La traîne des abonnements sur YouTube étant particulièrement longue, nous avons décidé de critères pour limiter le corpus : les chaînes animées par des personnes représentant leur propre opinion doivent atteindre au moins 500 abonnées pour être intégrées au corpus, être actives durant la période de collecte (les 6 derniers mois) et contribuer à des discussions relatives aux débats publics. Les chaînes de moins de 500 abonnées intégrées au corpus correspondent à des institutions de grande notoriété. Bien que suscitant très peu d'intérêt sur YouTube, leur intégration au corpus est déterminante pour comprendre les configurations renouvelées de l'espace public via YouTube.

Bien qu'imparfaite, incomplète et discutable, la méthodologie utilisée pour construire ce corpus met en évidence les multiples acteurs en compétition et, dans le même temps, représente les différents types de chaînes publiant du contenu à caractère politique et médiatique en France sur YouTube. Si des chaînes importantes et influentes peuvent être absentes du corpus et si des engagements politiques peuvent ne pas être représentés, cette liste permet néanmoins d'avoir un aperçu global de la structure du système YouTube. De plus, il serait illusoire de chercher à obtenir une cartographie exhaustive de l'ensemble des chaînes produisant des informations à caractère médiatique et politique ; il est aussi impossible de dresser la liste de toutes les entités économiques et politiques françaises, et de vérifier leur existence sur YouTube Il faut plutôt envisager cette méthode par liste de chaînes comme le moyen d'avoir une prise approximative sur le flux de contenus produits et diffusés sur la plateforme en matière médiatique et politique. La liste que nous avons dressée est un proxy toujours modifiable au gré de l'évolution des questionnements de recherche et de collaboration de recherche futur. Le caractère gigantesque de la plateforme invite à une approche « small data » qui, si elle est imparfaite, à au moins le mérite d'être intelligible.

Ainsi, il faut interpréter notre corpus comme un échantillon non aléatoire et raisonné de tous types de chaînes qui diffusent du contenu à caractère médiatique et politique (problèmes sociaux, moraux, scientifiques, économiques, écologiques, historiques etc.) Notre corpus permet de représenter YouTube comme un espace de lutte discursive autour de la fixation de l'agenda médiatique. A première vue, les acteurs s'affrontent dans cette lutte à armes égales : YouTube met tous les comptes sur un même plan, peu importe la nature du contenu produit ou diffusé, tous les comptes sont des chaînes. Peu importe les modalités

d’usage de la plateforme (faire défiler le fil de la page d’accueil de l’application, requêter tel ou tel sujet, consulter via l’alerte YouTube les dernières publications des chaînes auxquelles on est abonné, ou s’en remettre entièrement à la recommandation algorithmique), n’importe quel type de format peut potentiellement être rendu visible et lié à des enjeux d’actualité (on peut par exemple s’informer tout autant sur la célébration du 14 juillet sur la chaîne du gouvernement que sur celle d’un vlogueur historien amateur qui fait une vidéo d’actualité visant à décrypter les origines de cette célébration). Autrement dit, une étude de l’écosystème médiatique et politique sur YouTube doit mêler médias traditionnels, activistes, militants, organisations, vlogueurs anonymes – tous engagés dans une lutte discursive sur la manière de dire la réalité.

Produire des métadonnées sur les chaînes

Pour rendre compte des chaînes qui composent l’espace public et médiatique sur YouTube, nous dégageons des catégories de chaînes, en cherchant des régularités ou des regroupements de chaînes partageant des caractéristiques communes. C’est une manière de catégoriser des énonciateurs, en procédant par induction à partir de chaînes qui nous semblaient exemplaires de certaines formes d’expressivité sur YouTube. Cette méthode par induction implique de dégager des critères lors du visionnage. Si ces critères sont rarement univoques, il faut pouvoir justifier d’une certaine cohérence pour pouvoir agréger des chaînes par similarité. Une exigence de cohérence absolue mènerait à l’accumulation de partitions du corpus, ce qui aurait conduit d’une part à une multiplication des chaînes frontières susceptibles à elles seules de former des catégories et d’autre part, à un accroissement de catégories aux effectifs excessivement faibles étant donnée la petite taille de cette liste. Cette méthode d'annotation manuelle engage nécessairement la subjectivité du codeur, c'est pourquoi nous avons largement débattu collectivement de notre partition en 7 groupes. Le tableau présente chaque catégorie ainsi que les chaînes les plus prototypiques de chaque classe (la liste complète de chaînes relevant de chaque catégorie est consultable en annexe).

Décrivons brièvement ces catégories. Sur YouTube, la plupart des grands médias professionnels ont désormais une chaîne qui délivrent un contenu d’information grand public. C’est pourquoi nous avons créé une catégorie « Médias mainstream ». Dans cette même lignée de médias professionnels, nous avons distingué les chaînes d’« Informations locales » associées à un territoire géographique délimité (villes, régions). La catégorie « Politique » regroupe des chaînes qui ont en commun d’être explicitement soutenues par un mouvement politique. Ce sont les chaînes des partis politiques ou de leurs représentants. Nous avons aussi créé une catégorie « Gilets jaunes » afin de mieux comprendre la manière dont les Gilets jaunes participent eux-mêmes à la production médiatique. On trouve dans cette catégorie non seulement des chaînes créées par et pour les Gilets jaunes eux-mêmes, mais aussi toutes celles qui ont réorienté leur contenu en faveur du mouvement comme par exemple la chaîne Isadora Duncan (le journaliste dit des Gilets jaunes) ou bien encore celle des vlogueurs Verdi et Demos Kratos. La catégorie « Contre-information » est composée de chaînes dont la ligne éditoriale a pour ambition de réinformer son auditoire en révélant des informations que des personnes, réseaux ou groupes puissants tenteraient de dissimuler avec la complicité des médias professionnels. Parmi ces chaînes, on retrouve un nombre important de vlogs édités par des acteurs dont on retrouve les notices biographiques sur le

site wiki dissidence : J'suis pas content TV, Alain Soral, Pierre Jovanovic, Dieudonné etc. Nous avons également fait le choix d'ajouter à cette catégorie des médias professionnels comme TV Libertés ou Le Média pour tous dont les accointances avec les milieux d'extrême droite sont évidentes, mais non revendiqués, voire dissimulées. Pour l'essentiel, les contenus produits par ce type de chaîne sont régulièrement dénoncés par les médias traditionnels comme des chaînes de désinformation diffusant des discours complotistes et haineux. Une cinquième catégorie, « Médiation », indique des chaînes dont la ligne éditoriale est plus pédagogique qu'engagée, par exemple Xerfi canal ou Hugo Décrypte. Une sixième catégorie que nous avons nommée « Médias alternatifs et analyse » a permis de rassembler des chaînes spécialisées dans le décryptage de l'actualité (Osons causer, par exemple), la vulgarisation politique (L'aile à Stick, par exemple) et le genre reportage d'investigation (Le Media ou Mediapart, par exemple). Notons que certaines chaînes appartenant à des universités ont été intégrées à cette catégorie. Ces chaînes qui ont en commun l'analyse sont souvent attachées à une critique sociale, pouvant ainsi adopter le registre de la dénonciation.

Catégories	Description	Chaines exemplaires
Politique	Chaînes dont l'engagement politique est directement lié à un parti politique. La chaîne peut représenter une personne ou un parti.	- Jean-Luc Mélenchon - Groupe Républicains Assemblée nationale - LREM
Contre-information	Chaînes mettant l'accent sur le rôle principal de la manipulation cachée par quelques personnes puissantes ou des groupes en particulier (les sionistes, les féministes, les gays etc.) et/ou entretenant la confusion quant à leurs appartenances politiques extrémistes.	- ERTV Rhône-Alpes - Boulevard Voltaire - RT France
Gilets jaunes	Chaînes spécialement créées à la faveur du mouvement des Gilets jaunes ou qui ont orienté leur ligne éditoriale en faveur du mouvement, se revendiquant par-là Gilets jaunes.	- Gilets Jaunes Commercy - Éveil Global Conscience Gilets Jaunes, - Isadora Duncan
Médiation	Chaînes spécialisées dans le décryptage, la vulgarisation, pouvant néanmoins être attachées à une critique sociale.	- Hugo Décrypte - Xerfi Canal - Osons causer
Médias alternatifs	Chaînes de médias d'opinion et d'analyse politiques de l'actualité.	- Mediapart - Le Média - Thinkerview
Médias mainstream	Chaînes des journalistes professionnels qui délivrent un contenu d'information grand public. On retrouve à la fois les médias télévisuels, radio, la presse écrite, les émissions de TV, mais aussi les *pure players*.	- France Inter - L'Obs - Arte - Brut
Information locale	Chaines d'information « locale », associées à un territoire géographique délimité (villes, régions)	- Télé Lyon Métropole - France 3 Pays de la Loire - Journal Citoyen Haute-Marne

Tableau 1 : Description des catégories de chaînes codées manuellement avec quelques exemples de chaînes exemplaires

Délimitation et préparation du corpus

Disposant au départ d'un corpus volumineux de près de 350000 vidéos extraites sur une période s'étendant de juin 2018 à juillet 2019, représentant la sphère politique et médiatique sur YouTube dans son ensemble, nous avons voulu restreindre ce corpus aux seules vidéos présentant un intérêt, direct ou indirect, quant à la problématique du traitement du mouvement des Gilets jaunes. De plus, nous avons réduit les vidéos de notre corpus aux seuls segments textuels pertinents dans cette perspective. Cette sélection, qui a mobilisé différentes techniques, tant statistiques que de traitement du langage naturel, s'est déroulée en cinq étapes.

Premièrement, nous avons sélectionné les vidéos dont la catégorie YouTube, ou celle de la chaîne de la vidéo, est « Actualités et politique » ou « People et blogs », plus celles pour lesquelles il y a mention explicite des Gilets jaunes dans le titre, la description, les tags ou les sous-titres. De plus, le sous-titrage doit exister et être assez long (300 octets, correspondant à une quarantaine de mots pleins). De 350000 vidéos, le volume est réduit à 168000 vidéos à l'issue de cette première étape (figure 1).

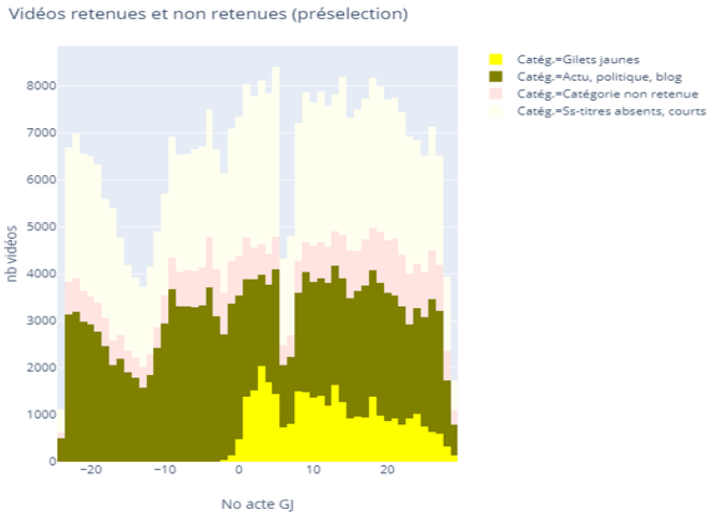

Figure 1 Vidéos retenues et non retenues dans la phase de pré-sélection dans la constitution du corpus. Vidéos éliminées en couleurs pales, vidéos conservées en couleurs vives.

Deuxièmement, nous avons cherché à affiner cette sélection en ne gardant que les vidéos dont les textes des sous-titres montrent une plus grande proximité avec la thématique des Gilets jaunes. En préalable, une analyse linguistique a été menée pour ramener les mots à leur forme lemmatisée. En effet, il nous a fallu préparer les données textuelles afin de les rendre manipulables pour les calculs. Dans l'écrasante majorité des cas, les sous-titres de vidéo sont générés à partir du logiciel de speech2text (reconnaissance vocale) de Google (texte fourni via l'API de YouTube), avec beaucoup d'erreurs d'interprétations qui vont au-delà des fautes orthographiques et se font souvent au niveau du découpage même de la chaine sonore selon les unités lexicales signifiées. Par ailleurs aucune marque de ponctuation ne permet d'isoler les phrases et autres propositions. Nous avons utilisé la bibliothèque spaCy pour conduire l'analyse linguistique des sous-titres de vidéo, mais en gardant à l'esprit que dans ces conditions, celle-ci est forcément souvent prise au dépourvu. Néanmoins, nous avons procédé aux phases de traitement classiques : élimination des incises hors-texte, signalées dans le sous-titrage par une chaîne de caractères entre crochets comme [Musique], [Applaudissements] (normalisation) ; détermination de la langue utilisée (les vidéos non en français sont éliminées) ; tokenisation ; détermination des catégories morpho-syntaxiques (*part-of-speech*) ; et lemmatisation. Par ce traitement, les sous-titres sont ainsi convertis en une liste de lemmes, auxquels sont rattachées leurs catégories morpho-syntaxiques. En vue de l'analyse thématique qui repose sur les sacs de mots (de termes), les mots les plus courants (*stop-words*), et appartenant aux catégories morpho-syntaxiques dites vides sont également éliminés, alors que les catégories dites pleines, noms, adjectifs, verbes, adverbes, numéraux et interjections sont conservées. Ce travail effectué, nous pouvons lancer la sélection des vidéos d'intérêt qui se base sur la similitude avec un texte de référence traitant des Gilets jaunes. Ce texte a été constitué d'extraits de l'article de Wikipedia France sur les Gilets jaunes, et complété par un florilège de citations sur ces événements, disponible sur un site aussi lié à Wikipedia. De façon

classique, on a constitué une matrice documents X termes, les documents étant les sous-titres des vidéos préalablement retenues (plus le texte de référence) et les termes étant les lemmes issus des analyses et sélections précédentes. Une transformation tf-idf des comptages de terme a été ensuite opérée sur cette matrice. Les vidéos sélectionnés sont alors celles pour lesquelles la similarité cosinus avec le texte de référence est la plus élevée, en retenant les 10 % les mieux classées, sélection qu'on a complété par les vidéos contenant explicitement l'expression « gilet jaune » dans le titre, la description, les tags ou les sous-titre, mais en ne retenant parmi ces dernières que les 80 % les plus similaires. Le résultat de cette sélection ramène le corpus à 33000 vidéos (figure 2).

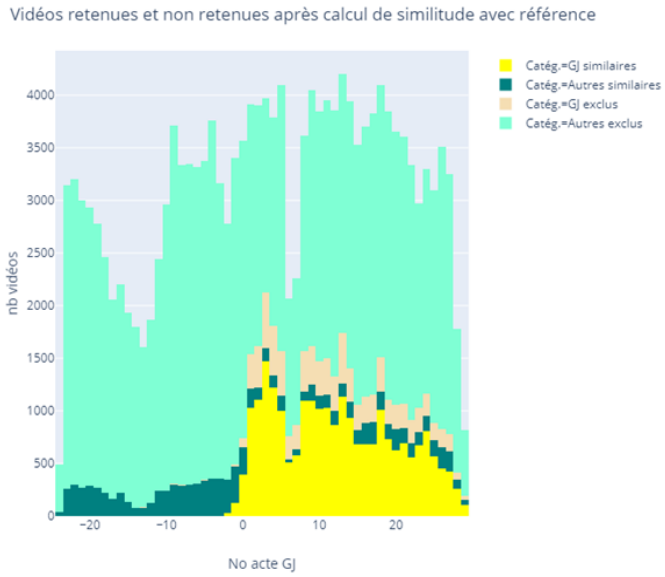

Figure 2 Vidéos retenues et non retenues après calcul de similitude avec référence. Vidéos éliminées en couleurs pales, vidéos conservées en couleurs vives.

La troisième étape de sélection vise à déterminer les groupes nominaux les plus significatifs, pour une analyse ultérieure plus précise. Les unités lexicales ne se réduisent pas aux unités morpho-syntaxiques que sont les mots, elles font le plus souvent intervenir des groupes de mots dont le sens n'est pas la simple composition des sens de leurs composants. Les groupes nominaux figés en sont en français les principaux représentants, qu'ils soient composés d'un nom et d'un adjectif (« gilet jaune ») ou de deux noms, en général reliés par la préposition « de » éventuellement associée à un article le plus souvent défini (« de la », « du », « des », par exemple « président de la république »). Ces groupes nominaux peuvent s'étendre sur plus de deux lexèmes nom ou adjectif (« mouvement des gilets jaunes »). Nous avons utilisé deux méthodes complémentaires pour établir ces groupes nominaux : d'une part, les suites (sans mentionner ici les prépositions et articles éventuels) (Nom, Nom), (Adjectif, Nom), (Nom, Adjectif) dans les catégories morpho-syntaxiques repérées par l'analyseur linguistique de spaCy et d'autre part une analyse statistique des bigrammes dont les associations sont les plus significatives dans le corpus retenu, et qui sont mesurées par une quantité dérivée de l'information mutuelle portée par les deux termes. Cette analyse s'est faite sur les lemmes retenus (donc sans tenir compte des stop-words comme les prépositions et articles). Les combinaisons (trigrammes) de trois lemmes ont été découvertes en faisant une nouvelle passe d'analyse statistique, cette fois-ci sur les suites de lemmes simples et bigrammes découverts lors de la phase précédente. L'examen manuel de la liste des bigrammes et trigrammes les plus fréquents a permis en outre de repérer grâce aux contextes ainsi fournis les mauvaises interprétations les plus fréquentes

du logiciel utilisé pour le speech2text de Google YouTube, d'y associer les corrections correspondantes, et de réinjecter ces corrections dans les textes devenus listes de lemmes simples, bigrammes et trigrammes. Ces textes peuvent alors être utilisés pour l'analyse thématique.

Quatrièmement, nous avons procédé à une réduction supplémentaire du nombre de vidéos en ne conservant que celles pour lesquelles un score combinant la proximité avec le texte de référence et la présence de topics liés au mouvement des gilets jaunes est suffisamment élevé. Autrement-dit, il s'agit de réduire le corpus en mobilisant un topic model et en ne conservant que les documents les plus pertinents selon les topics qui y sont présents. Pour ce faire, on a procédé à une analyse thématique par factorisation non-négative (NMF) de la matrice tf-idf des documents X termes, les termes étant les lemmes, bigrammes et trigrammes constitués à l'étape précédente. Nous avons retenu 64 topics pour cette opération, puis les avons classés manuellement après observation des nuages de mots correspondants, en distinguant : les topics de fond directement reliés au mouvement des Gilets jaunes ; les topics de fond indirectement reliés aux GJ (mais potentiellement intéressants pour l'étude) ; les topics de fond sans rapport aucun avec les GJ ; les topics de forme (vocabulaire le plus général, expressions parlées…). Pour la construction du score thématique, un coefficient allant de 1 à -1 a été affecté à chaque topic selon sa proximité avec la thématique générale des gilets jaunes, d'où un score thématique en combinant ce coefficient avec les coefficients de proportion des topics dans les documents. Un score combiné (2/3 score thématique et 1/3 score de similitude avec la référence) a été ensuite utilisé pour sélectionner les vidéos, en étant plus sévère pour les vidéos de la période avant la préparation et le mouvement des Gilets jaunes. Le nombre final de vidéos prises en compte se monte à un peu plus de 26000 (figure 3).

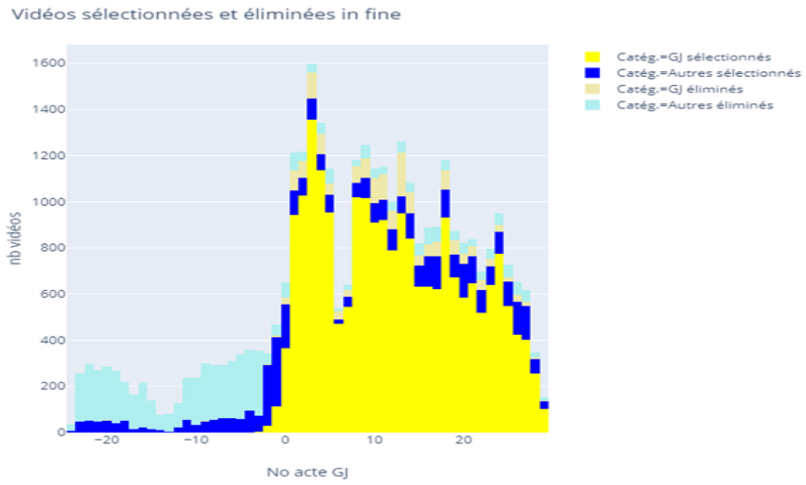

Figure 3 Vidéos sélectionnées et éliminées in fine. Vidéos éliminées en couleurs pales, vidéos conservées en couleurs vives.

Enfin, nous avons éliminé au sein des textes des sous-titres des vidéos retenus, des blocs de texte portant des thématiques spécifiques trop éloignées de celles des gilets jaunes. Ainsi, nous avons réduit les textes aux seuls passages qui sont en ligne avec le sujet d'étude, celui des Gilets jaunes, en se fondant sur l'évolution du contenu en topics au sein

de chaque texte de sous-titres. Pour ce faire, nous avons découpé les textes en blocs de 50 unités se chevauchant par moitié, et utilisant le modèle de topics constitué auparavant, avons obtenu la distribution des topics sur chacun de ces blocs, pour éliminer les blocs à forte proportion de topics de fond sans rapport aucun avec les Gilets jaunes et ceux dont les topics de fond en rapport avec les Gilets jaunes ont un poids extrêmement faible. Au bout du compte, seules quelques 7000+ vidéos n'ont subi aucune réduction de texte.

Méthode

Pour explorer ce corpus, nous nous sommes tournés vers une méthode d'analyse inductive des données linguistiques : les *topic models*, des algorithmes capables d'opérer des calculs sur une représentation numérique des documents textuels sans exiger de modèle a priori (Blei, 2012). Au lieu d'établir a posteriori une catégorisation des contenus, l'analyste se laisse en quelque sorte guider par l'identification algorithmique de classes latentes d'occurrence des termes lexicaux (mots, groupes de mots) dans les documents. Le calcul des *topics models* repose sur la distribution des termes lexicaux dans la collection des documents. Il produit des listes de *topics* constitués de termes qui co-occurrent dans les documents selon différents *patterns*. On peut considérer un topic comme un sujet (ou thème) abordé dans les documents, mais il faut bien voir aussi que certains de ces topics sont plus des marqueurs stylistiques que des sujets de discussion, d'où l'emploi du terme neutre (en français) de « topic ». Chaque document peut contenir plusieurs topics, et chaque topic a une cohérence interne qui dépend du niveau de coprésence des termes les plus représentatifs du topic dans les documents et par rapport aux autres topics. Après avoir présenté l'esprit de l'analyse de topics, nous rentrerons plus en détail dans les techniques de calcul, puis nous exposerons les résultats que nous avons obtenus.

L'esprit des topic models

Une fois les topics calculés, le travail de l'analyste consiste à déterminer ce à quoi cette cohérence se réfère, et donc comment le topic peut être interprété. C'est cette interprétabilité qui détermine si la méthode des topic models est utile ou non aux chercheurs en sciences sociales. Mais cette interprétabilité est difficile à définir car les clusters de mots générés à l'aide des topic models ne peuvent être utilisés pour répondre directement à des questions de recherche théorique en science sociale. De plus, les topics dégagés correspondent rarement à une catégorisation homogène, ce qui rend impossible la production automatique d'une taxonomie. En effet, la liste des topics obtenus s'assimile à une multitude de manière de faire communiquer les mots et les choses : des thèmes, des styles, des controverses, des événements, des cadrages etc. Ainsi, utiliser les topic models n'est pas une manière d'éviter le travail laborieux de codage manuel des contenus, comme on le croit trop souvent. C'est plutôt chercher l'étonnement dans la classification automatique des mots qui composent les textes et accepter un ordre des choses hétéroclite du point de vue de l'enquête sociologique.

Pour illustrer la nature hétérogène des topics générés et les difficultés d'interprétation qu'ils présentent, examinons l'exemple d'une vidéo de notre corpus.

> Topic 4 Taxation carburant (0.26017630179460366)
> Topic 56 Sentiment (0.24549021026076878)
> Topic 63 Pouvoir d'achat (0.2316606283404634)
>
> Chaîne : BFMTV
> Titre : Prix au sommet ? Plus chers que chez nos voisins ? Le vrai du faux sur la hausse du carburant
> Date : 13/11/2018
>
> s'il y a bien un élément qui cristallise la colère des gilets jaunes c'est la taxe sur le carburant qui ont tous les jours qu'un jour les jours c'est une orbite en cette taxe est-elle vraiment responsable de la hausse des prix oui et non la majeure partie de l'augmentation est en réalité due à celle du baril de pétrole en deux ans son coût est passé de 43 à 74 dollars ainsi le coût hors taxes d'un litre de gazole est passé de 44 à 65 centimes soit 21 centimes d'augmentation l'augmentation des taxes et de 7,6 centimes elles ne représentent donc qu'un tiers de la hausse des prix mais le prix du carburant a tâté l'aujourd'hui des sommets non en termes de pouvoir d'achat en 1981 un litre d'essence coûte et 52 centimes mais le smic horaire était de 2,18 euros pour une heure travaillée on pouvait donc s'acheter trois litres d'essence c'est deux fois moins qu'aujourd'hui et par rapport à ses voisins européens la france est tellement bien lotis non en france le prix du gazole est en moyenne de 1,51 euro c'est l'un des tarifs les plus élevés mais dans plusieurs autres pays les consommateurs payent plus cher c'est le cas de l'italie la belgique la grande-bretagne et la suède dans la plupart des pays européens les taxes sur le carburant sont d'ailleurs équivalente à celle de la france représentant plus de la moitié du prix à la pompe

Cette vidéo est associée à des degrés divers à plusieurs topics (le score de probabilité indiqué entre parenthèse à côté des noms de topic dans l'exemple ci-dessus). Les termes peuvent aussi être associés à des degrés divers à différents topics (les mots "travaillée" et "hausse", bicolores). Notons au passage, mais nous ne pouvons pas en rendre compte ici à partir d'un seul document, que le même terme dans deux documents différents peut être associé à des topics différents.

Etant donné le contenu de la vidéo, de nombreux mots ou expressions renvoient à la taxation du carburant, indiqués en bleu. Ce topic comprend des mots tels que "taxe", mais aussi des expressions (groupes nominaux) comme "taxe sur le carburant, "prix à la pompe" et "augmentation des taxes". L'autre topic principal, le pouvoir d'achat, est indiqué en vert et contient des mots tels que "pouvoir d'achat", "smic horaire" et "hausse". Enfin, un certain nombre de termes tels que "colère", "travaillée" et "pays" sont tirés du topic des sentiments. D'autres mots renvoient à d'autres topics, mais nous ne les soulignons pas.

Cet exemple met en évidence un certain nombre de points intéressants concernant la méthode des topics model. Tout d'abord, le fait que les topics qui en résultent ont un sens concret et correspondent à la façon dont nous définirions les thèmes de la vidéo montre que, dans une certaine mesure, la notion qu'un humain se fait d'un topic correspond aux classes latentes identifiés par l'algorithme. Ensuite, aucun codage a priori n'a été utilisé par l'algorithme, de sorte que les topics de ce document ont été trouvés de manière totalement automatique. Enfin, ce document est présenté en trois thèmes principaux. Dans un système de codage qui oblige à avoir un seul sujet par document, il serait très difficile de choisir le thème « dominant » pour cette vidéo. La coprésence des topics tels que taxation du carburant, pouvoir d'achat et sentiment peut être interprétée comme une façon de cadrer le problème des Gilets jaunes.

Une pluralité de manière de modéliser les topics

Si l'esprit de la méthode des topics models est relativement simple à cerner, les techniques de calculs sont plus complexes à comprendre tant par la sophistication des algorithmes que par la pluralité des approches disponibles.

Le résultat d'une analyse thématique se présente sous la forme de deux matrices : d'une part, une matrice documents X topics, représentant sur chaque ligne la distribution des topics sur le document correspondant et d'autre part, une matrice topics X termes, représentant sur chaque ligne les poids des différents termes lexicaux (mots ou expressions) dans le topic correspondant (et dans chaque colonne le degré d'appartenance au topic du terme correspondant). Un document peut donc porter sur plusieurs topics, et un même terme être employé par plusieurs topics, bien sûr avec des poids différents d'un topic à l'autre.

On distingue deux grandes catégories d'approche algorithmique pour l'analyse : les approches probabilistes générativistes, où chaque document est considéré comme généré à partir de probabilités de topics sur les documents, et de termes sur les topics, chacune avec leur propre distribution ; et les approches issues de l'algèbre linéaire, en procédant par factorisation de matrices. Dans les deux approches, on part d'une matrice documents X termes, ou à chaque combinaison document X terme on associe un coefficient marquant l'importance du terme dans le document. On est donc dans une approche dite « sacs de mots » où les positions respectives des mots (termes) les uns par rapport aux autres, à savoir donc l'influence de la syntaxe et à un niveau plus élevé, la structuration du discours sont supposés avoir une influence restreinte par rapport à la tâche qui nous intéresse ici.

Historiquement, après quelques précurseurs, la première technique d'application répandue a été une factorisation de matrice (par SVD, Singular Value Decomposition, avec la méthode LSA, Latent Semantic Analysis, Deerwester et al., 1990). Elle a été suivie d'une première technique probabiliste, utilisant des distributions conditionnelles des topics sur les documents, et des topics sur les termes (pLSA, Probabilistic Latent Semantic Analysis, Hofmann, 1999). pLSA a elle-même été raffinée par la méthode LDA (LDA, Latent Dirichlet Allocation, Blei et al., 2003), qui est très rapidement devenue la méthode standard par défaut pour l'analyse thématique, et qui a donné lieu à un certain nombre d'adaptations et d'enrichissement, les plus notables étant : CTM (Correlated Topic Model, Blei et al., 2007), une évolution de LDA qui cherche à exploiter les corrélations entre topics au sein des documents ; DTM (Dynamic Topic Model, Blei et al., 2006), un enrichissement de LDA où les topics évoluent entre les différentes périodes temporelles ; STM (Structural Topic Modeling, Roberts et al., 2014, Lindstedt, 2019), une extension de LDA qui y adjoint les influences de facteurs covariants via un modèle linéaire généralisé ; et à partir de 2016 diverses intégrations exploitant les plongements de mots (word2vec, Glove pour commencer, par exemple, Dieng et al., 2019). Mais force est de constater qu'en 2020, aucune de ces adaptations n'a réellement détrôné LDA en tant que standard par défaut.

Parallèlement, une amélioration notable dans l'approche basée sur la factorisation de matrices a été apportée par NMF (Non-negative Matrix Factorization, Lee et al., 1999, Arora et al. 2012) qui permet une factorisation approchée de la matrice documents X termes en deux matrices uniquement composées de coefficients positifs ou nuls, et donc aisément

interprétables en tant que coefficients associés à des topics. Plusieurs modalités de calcul en ont été développées, certaines plus stables que d'autres.

Les complexités de toutes ces méthodes tiennent essentiellement aux algorithmes nécessaires pour la découverte des distributions probabilistes (ou de la factorisation la plus efficace selon le cas) et autres initialisations une fois les hyperparamètres fixés. A noter que quelle que soit la technique utilisée, un paramètre est toujours à fixer, c'est le nombre de topics K, dont il faut ensuite estimer la pertinence après exécution par rapport à d'autres valeurs (et sans compter les combinaisons d'autres hyperparamètres à considérer par ailleurs).

Un autre point important est celui de la matrice documents X termes à utiliser comme point de départ, et notamment le traitement préliminaire des termes les plus et les moins fréquents. Les préconisations généralement suivies sont pour les méthodes probabilistes de partir des comptages ordinaires, mais en éliminant les termes les moins et les plus fréquents (selon leur fréquence totale ou selon le nombre de documents dans lesquels ils apparaissent). En effet ces méthodes sont censées partir de modélisations de distributions de probabilités des termes sur les topics, puis les documents. Au contraire, pour la méthode NMF, il est préconisé de partir de coefficients déterminés par la pondération tf-idf pour réduire les poids des termes apparaissant un peu partout dans les documents (voir par exemple Green et al., 2014) (ce qui n'empêche pas non plus d'éliminer les termes trop peu fréquents).

Dans cette étude nous avons choisi de mettre en œuvre les méthodes suivantes, en nous fondant sur la disponibilité et la popularité des implémentations disponibles (librairies Python et package R) : NMF (librairie Python scikit-learn) ; LDA (librairie Python gensim) ; STM (package R stm). Plusieurs méthodes existent pour évaluer les résultats (les topics) d'une analyse thématique. Elles ont d'abord été conçues et utilisées pour déterminer le nombre K de topics optimal pour l'analyse, en général dans le cadre d'une et une seule méthode (LDA, etc.). Mais elles peuvent tout aussi bien être utilisées pour comparer entre eux les résultats issus de différentes méthodes (typiquement NMF vs LDA).

On peut d'une part distinguer entre méthodes quantitatives (métriques) et méthodes qualitatives et d'autre part distinguer entre les méthodes qui s'intéressent à la distribution des topics sur les documents, et celles qui s'intéressent à la distribution des termes sur les topics. Ces méthodes peuvent enfin s'intéresser aux topics pris un par un (intra-topic) ou aux topics comparés entre eux (inter-topics).

Les mesures quantitatives de distribution des topics sur les documents sont des métriques de cohérence sémantique : elles se calculent au niveau de chaque topic (intra-topic), et peuvent ensuite être moyennées sur l'ensemble des K topics dégagés par l'analyse. Au niveau d'un topic donné, on se fonde sur les N*(N-1)/2 co-occurrences deux par deux des N termes les plus représentatifs du thème ; plus ces termes apparaissent ensemble, plus la cohérence sémantique est importante. Il existe plusieurs métriques de cohérence sémantique (Mimno et al., 2011, Röder er al., 2015), dont LCP (Log Conditional Probability, encore appelée UMass) et NPMI (Normalized Pointwise Mutual Information). Le calcul originel de LCP s'appuie sur les co-occurrences au sein des documents (et est plus rapide à

calculer), alors que NPMI prend en compte une fenêtre glissante de taille W termes (et est plus précise).

Les mesures quantitatives de distribution des termes sur les topics sont multiples. On trouve des métriques de cohérence par rapport à un espace vectoriel des termes : elles se calculent au niveau de chaque topic (intra-topic) en considérant deux à deux les proximités sémantiques des N premiers termes du topic dans l'espace vectoriel (similarité de cosinus), et les résultats des K topics peuvent ensuite être moyennés. Cet espace vectoriel a été d'abord un espace statique issu des méthodes de plongement de mots de 1ère génération (word2vec, Glove, etc.), mais on peut imaginer utiliser aussi des plongements contextuels (BERT…). On trouve aussi des métriques d'exclusivité : au niveau de chaque topic (intra-topic), pour chacun des N premiers termes du topic, on considère son poids dans le topic relativement à son poids dans tous les topics. La mesure FREX (Frequency-Exclusivity) (Roberts et al., 2014), est la plus connue de ces métriques et est la moyenne harmonique sur les termes entre une partie exclusivité et une partie fréquence (et une pondération favorisant l'exclusivité). Enfin, on peut aussi calculer des métriques (d'absence de) généralité : Ce sont des métriques inter-topics. On peut par exemple considérer les rapports entre intersection et union des N premier termes de chaque ensemble de topics pris deux par deux (mesure de Jaccard), et moyenner cette mesure sur les K*(K-1)/2 combinaisons de topics. On peut aussi s'intéresser à la distribution du nombre de termes rencontrés dans X topics (au niveau des N tops termes de chaque topic) (Green et al., 2014).

A ces deux séries d'évaluations quantitatives des topics, il est fort recommandé d'ajouter une évaluation qualitative de distribution des termes sur les topics ou des topics sur les documents. La première se fonde sur l'examen des listes de N premiers termes de chaque topic, ou mieux par examen des nuages de mots représentant ces topics. La seconde vise à compléter l'examen des nuages de mots par un retour au texte des documents les plus représentatifs des topics que l'on souhaite valider plus précisément[1].

Comment choisir la bonne méthode et le bon nombre de topics ?

Dans le cadre de cette étude, les résultats en ont été évalués qualitativement par examen des nuages de mots et des documents représentatifs des topics, et quantitativement en utilisant les métriques et évaluations suivantes : cohérence sémantique (LCP) ; cohérence (notée ici W2V) par rapport à espace vectoriel word2vec généré sur le corpus (via la fonction correspondante de la librairie gensim) et la métrique d'exclusivité (FREX). Nous pouvons porter graphiquement l'évolution des métriques de qualité selon le nombre de topics (de 8 en 8, entre 8 et 80), avec chaque méthode (NMF, LDA, STM) en parallèle :

[1] Il est à noter que les packages standard n'implémentent qu'incomplètement l'ensemble de ces mesures, et que celles-ci sont à redévelopper là où elles manquent, ce qui permet aussi d'avoir un cadre cohérent de comparaison pour les résultats de deux méthodes différentes. Pour mémoire, la librairie LDA de gensim implémente les cohérences sémantiques LCP et NPMI, et de plus la métrique c_v, plus sophistiquée et efficace que les deux précédentes. La librairie scikit-learn de NMF n'offre qu'une mesure de distance entre le produit de la factorisation et la matrice documents X termes d'origine, et le package STM offre les deux métriques LCP et FREX.

Figure 4 Comparaison pour les trois méthodes de topic model des valeurs des métriques de cohérences sémantiques (LCP et W2V) et de fréquence et d'exclusivité (FREX), basées pour chaque topic sur les 10 principaux termes (à gauche) et les 20 principaux (à droite).

Il apparaît à l'examen des valeurs de la métrique FREX que la distribution des termes sur les topics est plus étroite (plus exclusive) suite à la factorisation NMF qu'elle ne l'est sur les topics issus des méthodes probabilistes LDA et STM ; cela se traduit aussi à l'examen visuel des nuages de mots (de termes) où peu de termes se rencontrent éparpillés sur plusieurs topics différents. De façon analogue, les nuages de mots semblent sémantiquement beaucoup plus cohérents avec la méthode NMF qu'ils ne le sont avec LDA et STM, ce qui est confirmé à l'examen de la métrique W2V. Un point à souligner est que Blei (l'auteur de la méthode LDA) est bien conscient d'une trop grande généralité des termes qui se retrouvent hautement positionnés sur plusieurs topics, et propose une mesure alternative des poids des termes dans les topics, mesure qui applique une pondération sur les coefficients quelque peu analogue à la transformation tf-idf utilisée pour NMF, mais postérieurement à la découverte des topics et non antérieurement.

Par contre, la mesure de cohérence LCP sur les co-occurrences des termes au niveaux des documents pris chacun dans leur totalité, favorisent les mesures fondées sur les distributions de probabilité (LDA et STM). Une explication en est que les termes relativement généraux se trouvent dans une très grande partie des documents, et co-occurrent donc très souvent avec les autres termes. Lorsqu'on examine les topics isolément les uns des autres, on constate que les topics les plus spécifiques sont ceux qui présentent les moins bons

résultats selon cette métrique. Il serait bon d'examiner si cet effet est annulé lorsqu'on utilise des fenêtres glissantes pour mesurer les co-occurrences[2].

Il convient enfin de remarquer l'influence du nombre de termes utilisés par les calculs. En considérant 20 termes plutôt que 10, les cohérences vectorielles diminuent, ce qui est normal, les termes suivants introduisant plus de dispersion. Quant aux gros écarts constatés pour les deux autres mesures, ils sont un simple effet du nombre de termes considérés. Toutes ces constatations sont également en ligne avec les résultats de plusieurs études.

Une fois ces métriques calculée, la problématique classique est celle de la détermination du nombre idéal de topics au vu de l'évolution des diverses métriques de qualité, en trouvant le « sweet spot » parvenant au meilleur compromis entre les différentes exigences de qualité. Nous pouvons déjà retenir que quelle que soit la méthode utilisée, la cohérence sémantique LCP diminue avec le nombre de topics, alors que l'exclusivité au contraire augmente. Quant à la cohérence par rapport à l'espace vectoriel, elle ne présente pas une tendance bien nette. Nous reviendrons sur ce point ultérieurement, mais auparavant, il convient de se fixer sur la méthode à retenir.

Et pour ce, deux autres points importants, et liés entre eux, doivent être abordés : la stabilité des topics découverts, et la dispersion des topics selon le volume plus ou moins important qu'ils occupent sur l'ensemble des documents. Les topics construits par une analyse thématique peuvent ne pas se retrouver, ou être déformés dans une autre analyse thématique ultérieure, pour peu que des petites perturbations soient apportées au corpus (quelques documents en plus ou en moins), que l'on fasse varier légèrement (un ou deux en plus ou en moins) le nombre de topics à déterminer, ou même que l'on change l'initialisation aléatoire du calcul. Selon la littérature, NMF est une méthode plus stable que les deux autres.

Certaines méthodes, dès lors que le nombre de topics est relativement élevé, créent des topics relativement peu stables, qui ne se retrouvent pas d'une exécution à l'autre, et dont la masse globale sur l'ensemble des documents est très faible. Ceci est surtout vrai de la méthode LDA, comme cela peut être constaté sur les visualisations suivantes (pyLDAvis), la surface des cercles y étant proportionnelle à la masse globale du topic.

[2] il convient de noter que changer les hyper-paramètres spécifiques aux méthodes utilisées ne change pas grand-chose à leurs positionnements relatifs vis-à-vis de ces métriques. Pour LDA, nous avons pris les paramètres par défaut de la librairie gensim, notamment pour α et η (β) qui commandent la concentration des distributions respectivement des topics sur les documents, et des termes sur les topics. Pour STM, nous avons pris comme co-variables les catégories de chaîne, les numéros d'acte GJ, et une interaction simplifiée entre les deux.

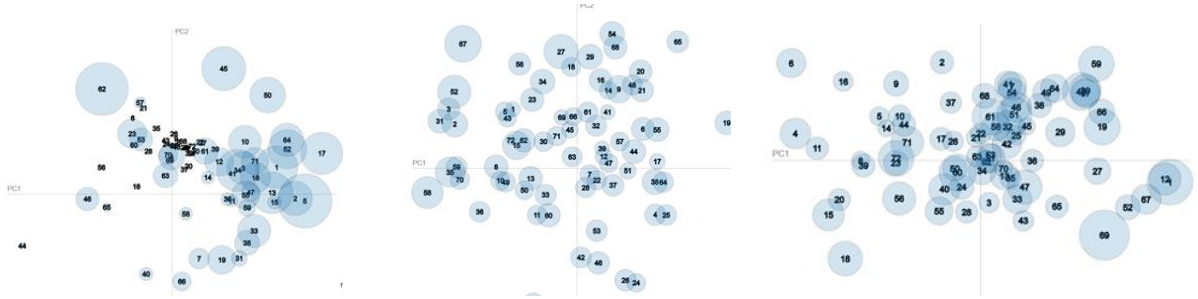

Figure 5 Projection des distances inter-topics pour K=72 calculée avec l'outil pyLDAvis. La taille des cercles correspond aux poids des topics.

On constate que NMF et STM sont beaucoup plus équilibrés à ce niveau. Dans l'ensemble, et sous réserve de calculs d'autres coefficients de cohérence sémantique (à baser sur des fenêtres de texte), et de l'application aux méthodes probabilistes de coefficients de pondération au sein de la matrice topics x termes, nous privilégions les résultats de la factorisation de matrices par NMF.

Reste à fixer le nombre de topics idéal, et plus précisément la constellation de topics la plus intéressante pour notre étude. En ramenant chaque métrique à occuper une plage entre 0 (valeur minimum prise sur l'ensemble des nombres de topics) et 1 (valeur maximum), nous pouvons mieux apprécier l'évolution parallèle des métriques selon le nombre de topics.

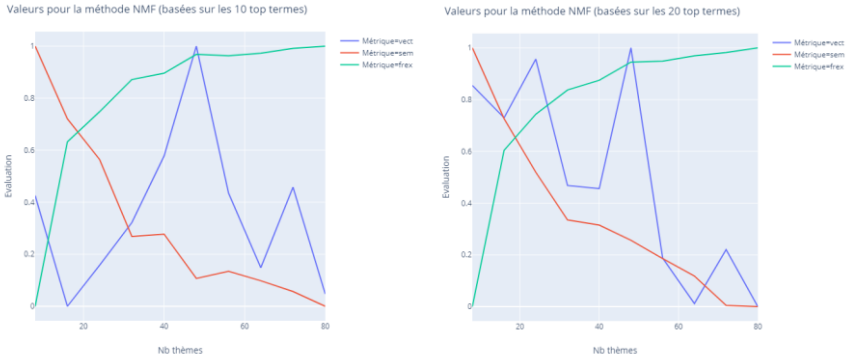

Figure 6 Valeurs des deux métriques de cohérence sémantique (LCP et W2V) et de la métrique de fréquence et d'exclusivité (FREX) pour la méthode NMF. A gauche, basé sur 10 termes, à droite sur 20 termes.

Les tendances de croissance de la métrique FREX, de décroissance de la métrique de cohérence sémantique LCP avec le nombre de topics se confirme sur ce graphe ; pour la métrique de cohérence vectorielle, la tendance semble varier selon le nombre de top termes retenus. Plus précisément, on semble atteindre un quasi-plateau autour de la cinquantaine de topics, tant pour la croissance de FREX que pour la décroissance de LCP, et éventuellement un maximum de la métrique de cohérence vectorielle autour de ce nombre de topics, si on se limite à un faible nombre de top termes.

Si un nombre autour de 50 serait à priori le nombre de topics idéal selon les métriques calculées, nous avons choisi d'en conserver 72 car l'évaluation visuelle a permis de constater qu'à ce niveau les topics demeuraient suffisamment cohérents et apportaient un niveau de granularité plus riche pour les analyses ultérieures. Il est courant dans l'évaluation

du nombre de topics que les évaluations humaines, donc visuelles, ne corroborent pas tout à fait celles des algorithmes.

La signification des 72 topics peut être rapidement appréhendée en listant pour chacun d'entre eux les 10 termes qui y sont les plus représentés :

topic 0 - action : faire, essayer, aller, sorte, besoin, pouvoir, déjà, passer, train, mettre
topic 1 - mesures : mesure, annoncer, ministre, annonce, président_république, prendre, édouard_philippe, réponse, milliards_euros, répondre
topic 2 - manifestations 1 : manifestant, forces_de_l'ordre, gaz_lacrymogène, affrontement, crs, disperser, place_de_la_république, calme, tension, situation
topic 3 - projet : projet, association, ici, travail, place, mettre, travailler, également, permettre, action
topic 4 - taxation carburants : taxe, voiture, diesel, carburant, essence, augmenter, véhicule, taxer, transition_écologique, taxe_carbone
topic 5 - discours politique : politique, chose, croire, moment, société, pouvoir, manière, évidemment, forme, finalement
topic 6 - grand débat 1 : débat, sujet, organiser, débattre, grand_débat_national, topic, lettre, participer, proposition, président_république
topic 7 - Macron !! : macron, macro, france, bon, monsieur_macron, merdia, castaner, sarkozy, voter, vouloir
topic 8 - droit & loi : loi, texte, droit, article, député, constitution, liberté, manifester, sénat, assemblée_nationale
topic 9 - partis politiques : gauche, droite, parti, extrême_droite, socialiste, parti_socialiste, benoît_hamon, républicain, libéral, génération
topic 10 - finances : argent, banque, système, payer, état, milliard, dette, riche, financier, économie
topic 11 - gilets jaunes : gilet_jaune, jaune, gilet, rond-point, rond_point, soutenir, eric_drouet, crise, rencontrer, début_du_mouvement
topic 12 - police 1 : policier, police, quartier, commissariat, flic, suicide, violences_policières, enquête, police_nationale, effectif
topic 13 - police 2 : collègue, hiérarchie, fonctionnaire, police_nationale, maintien_de_l'ordre, monsieur_le_ministre, suicide, ordre, sécurité, service
topic 14 - familier : oui, bon, savoir, vrai, aimer, sûr, petit, bonjour, accord, mal
topic 15 - écologie : écologie, climat, écologique, transition_écologique, action, écologiste, planète, environnement, climatique, europe
topic 16 - débat public : monsieur, monsieur_macron, avoir, venir, écouter, être, entendre, croire, savoir, demander
topic 17 - très familier : truc, ouais, mec, merde, putain, savoir, live, bon, coup, salut
topic 18 - discours action : falloir, croire, sûr, solution, problème, important, savoir, mettre, moment, arrêter
topic 19 - infos générales : matin, rtl, hier, bonjour, hier_soir, heure, strasbourg, ministre, édouard_philippe, midi
topic 20 - syndicat : syndicat, cgt, grève, premier_mai, syndical, convergence, salarié, travailleur, syndicaliste, cfdt
topic 21 - violences (condamnées) : violence, violent, condamner, voir, légitime, image, violences_policières, justifier, réponse, république
topic 22 - médias : journaliste, média, presse, information, journal, médiatique, image, bfm, france, article
topic 23 - en direct : demain, soir, bonsoir, heure, attendre, venir, journée, évidemment, monde, entendre
topic 24 - mobilisation : mobilisation, mobiliser, acte, semaine, continuer, rassemblement, chiffre, poursuivre, ministère_intérieur, week-end
topic 25 - antisémitisme : antisémitisme, juif, antisémite, haine, acte_antisémite, alain_finkielkraut, acte, antisionisme, sioniste, racisme
topic 26 - manifestations 2 : manifestation, manifester, déclarer, interdire, manif, liberté, organisateur, organiser, lieu, arrêter
topic 27 - blocages : bloquer, blocage, camion, ici, péage, rond-point, autoroute, rond_point, action, automobiliste
topic 28 - mouvement GJ : mouvement, soutenir, leader, soutien, france, parti_politique, organiser, structurer, continuer, mouvement_social
topic 29 - action gouvernement : gouvernement, ministre, politique, édouard_philippe, porte_parole, exécutif, opposition, majorité, mettre, benjamin_griveaux
topic 30 - revenus salaires : euro, 100, payer, monnaie, smic, prime, euro_mois, 200, toucher, mois
topic 31 - modalités : effectivement, justement, finalement, cas, peut-être, évidemment, également, vrai, rappeler, dire
topic 32 - à Paris : paris, parisien, ville, capitale, france, province, quartier, partout, arrondissement, bordeaux
topic 33 - féminisme : femme, homme, féministe, droit_des_femmes, victime, viol, combat, féminisme, mari, lutte
topic 34 - commerce & GJ : commerçant, commerce, noël, magasin, chiffre_d'affaires, centre_ville, client, boutique, fermer, ville
topic 35 - lycéens étudiants : lycéen, lycée, jeune, étudiant, réforme, élève, établissement, professeur, enseignant, université
topic 36 - retraites : retraite, retraité, réforme, travailler, an, pension, 62, fonctionnaire, réforme_des_retraites, fonction_publique

topic 37 - grand débat 2 : grand_débat, grand_débat_national, proposition, contribution, participer, réunion, réponse, grand, attendre, fin
topic 38 - agriculture : produit, agriculteur, prix, consommateur, producteur, agriculture, acheter, magasin, client, vendre
topic 39 - réseaux sociaux : vidéo, youtube, facebook, chaîne, commentaire, réseau_social, live, internet, twitter, message
topic 40 - fiscalité : impôt, payer, riche, taxe, fiscal, fiscalité, impôt_sur_revenu, revenu, baisser, milliard
topic 41 - peuple & pouvoir : peuple, pouvoir, peuple_français, révolution, peuple_de_france, démocratie, élite, pays, système, france
topic 42 - cortège manifestation : cortège, place, ici, heure, rue, place_de_la_république, rassemblement, rejoindre, calme, marche
topic 43 - affaire Benalla : elysée, alexandre_benalla, affaire, monsieur_benalla, sénat, passeport, benalla, mediapart, affaire_benalla, justice
topic 44 - Corse : corse, nationaliste, île, dialogue, bastia, région, continent, ajaccio, élu, visite
topic 45 - présidence république : emmanuel_macron, président, chef_d'état, président_république, crise, elysée, quinquennat, politique, communication, françois_hollande
topic 46 - général 1 : personne, vraiment, essayer, coup, passer, chose, monde, forcément, justement, rapport
topic 47 - général débats : question, poser, répondre, pose, réponse, sujet, savoir, cas, évidemment, exemple
topic 48 - démocratie : citoyen, démocratie, élu, démocratique, politique, proposition, pouvoir, institution, justement, référendum_initiative_citoyenne
topic 49 - violence policière 1 : blessé, arme, grenade, utiliser, maintien_de_l'ordre, blessure, oeil, lbd, blesser, flashball
topic 50 - élections européennes : liste, voter, européen, élections_européennes, rassemblement_national, élection, candidat, campagne, parti, europe
topic 51 - violence policière 2 : gazer, taper, retraité, police, aimer, crs, mal, manifeste, plaire, continuer
topic 52 - casseurs : casseur, black_block, casser, forces_de_l'ordre, maintien_de_l'ordre, christophe_castaner, ministre_intérieur, dispositif, interpeller, casse
topic 53 - RIC : référendum, constitution, voter, référendum_initiative_citoyenne, vote, suisse, démocratie, élection, parlement, référendum_initiative_populaire
topic 54 - géné. live YT : ami, live, constitution, petit, monde, aller, constituer, bon, salut, ici
topic 55 - familles : enfant, an, famille, parent, vie, école, vivre, jour, venir, mère
topic 56 - expression sentiments : colère, exprimer, entendre, comprendre, pays, croire, profond, dialogue, manifester, réponse
topic 57 - politique locale : maire, commune, élu, ville, habitant, territoire, département, mairie, bordeaux, métropole
topic 58 - nation française : français, france, président_république, pays, croire, républicain, voir, réalité, sujet, évidemment
topic 59 - général 2 : chose, vouloir, accord, dire, savoir, aller, avoir, passer, mettre, prendre
topic 60 - macro économie : pourcent, chiffre, sondage, baisse, croissance, an, 10, 20, 2018, hausse
topic 61 - 1er mai : hôpital, premier_mai, christophe_castaner, intrusion, patient, ministre_intérieur, médecin, infirmier, soignant, service_réanimation
topic 62 - pol. sécurité routière : radar, route, 80_kilomètres_heure, vitesse, sécurité_routière, automobiliste, accident, chiffre, département, voiture
topic 63 - pouvoir achat : pouvoir_d'achat, salaire, augmenter, smic, retraité, salarié, augmentation, prime_d'activité, revenu, travail
topic 64 - champs-elysées : champs_elysées, place_de_l'étoile, avenue, crs, arc_de_triomphe, image, champ, avenue_champs_elysées, ici, rassemblement
topic 65 - revendication : revendication, entendre, référendum_initiative_citoyenne, porter, justement, revendiquer, exprimer, pouvoir_d'achat, répondre, demande
topic 66 - actes GJ : samedi, manifester, semaine, week-end, appel, samedi_prochain, rue, vendredi, acte, appeler
topic 67 - France insoumise : jean_luc_mélenchon, france_insoumise, politique, perquisition, marine_le_pen, mélenchon, député, insoumis, rassemblement_national, assemblée_nationale
topic 68 - description manifestation : voir, vraiment, passer, regarder, ici, image, train, petit, aller, venir
topic 69 - boxeur : cagnotte, boxeur, christophe_dettinger, gendarme, soutien, leetchi, soutenir, forces_de_l'ordre, frapper, blessé
topic 70 - justice & GJ : justice, avocat, prison, procès, juge, tribunal, condamner, juger, dossier, judiciaire
topic 71 - armée : militaire, armée, guerre, arme, soldat, mort, pays, général, france, opération

Tableau 2 Liste des 72 topics.

Résultats

Calculer des topics n'est pas une fin en soi. Les topics sont des variables à partir desquelles d'autres calculs peuvent être envisagés. Les vidéos rassemblées en corpus constituent un

ensemble sur lequel nous possédons de nombreuses informations externes, notamment des métadonnées sur les chaînes et le temps.

Dans la littérature en sciences sociales mobilisant les *topics models*, les topics servent à des tests d'hypothèses destinés à vérifier des hypothèses a priori sur des relations entre les topics et des variables (DiMaggio et al., 2013; Tsur et al., 2015). Dans cet article, nous avons plutôt suivi une approche exploratoire pour rechercher et découvrir des relations entre les topics entre eux et des variables externes, mais sans hypothèse *a priori* (Tukey, 1977). Nous avons privilégié trois techniques analytiques nous permettant de produire du sens et d'*ordonner* les topics : l'arbre des divergences inter-topics, le réseau de similarité de topics et les profils de topics par catégorie de chaînes et par acte des manifestations.

Avant de commenter ce que ces trois techniques analytiques donnent à voir, il nous faut rappeler la posture analytique avec laquelle nous envisageons cette lecture du corpus. Nous nous inspirons ici des travaux de Franco Moretti sur l'étude quantitative de corpus littéraire : « Lorsque nous étudions 200 000 romans au lieu de 200, dit Moretti, nous ne nous contentons pas de faire la même chose à une échelle 1 000 fois plus importante : nous les étudions de façon différente. Ce changement d'échelle modifie notre relation avec notre objet d'étude, et modifie de fait *jusqu'à l'objet lui-même* (Moretti, 2016) ». Cette approche de la lecture distante implique de traiter les corpus comme des « objets artificiels », la réalité sociale s'observant « *au sein même des abstractions* » produites par les méthodes computationnelles.

Dans les pages qui suivent, il n'y aura pas de retour au texte, dans le sens d'une démonstration par la citation verbatim des contenus pour illustrer le sens des topics[3]. Bien sûr, nous avons consulté un nombre important de vidéos lorsque nous avons voulu comprendre les topics. Mais le sens du corpus que nous souhaitons produire passera seulement par les abstractions computationnelles, c'est-à-dire les topics et la série de visualisations permettant de les explorer, de découvrir des relations entre les variables et de rendre compte d'une signification d'ensemble du traitement médiatique des Gilets jaunes.

Nous allons procéder en deux temps. Tout d'abord le dendrogramme et le réseau des topics similaires nous servirons à comprendre la nature des topics : s'agit-il de sujet, de controverse, d'épisode, d'événement etc. Ensuite, nous répondrons au thème principal de cette enquête, à savoir l'analyse de l'agenda médiatique selon les catégories d'acteur et les 24 premiers actes du mouvement.

L'arbre des divergences inter-topics

La représentation sous forme d'arbre hiérarchique ou dendrogramme, produit à partir d'un algorithme de classification hiérarchique, constitue une première étape pour apporter une visualisation d'ensemble des topics. Comme le décrit Lebart, « les regroupements effectués à chaque pas de l'algorithme de classification hiérarchique rassemblent des éléments qui

[3] De notre point de vue, de nombreuses études en humanité numérique commettent l'erreur de traiter de manière computationnelle des corpus, mais en interprétant les résultats comme des objets qualitatifs.

sont plus ou moins proches entre eux [...] La représentation sous forme de dendrogramme d'une classification hiérarchique matérialise bien le fait que les classes formées au cours du processus de classification constituent une hiérarchie indicée de classes partiellement emboîtés les unes dans les autres. » (Lebart, 1994). Un arbre peut être considéré comme la description simplifiée d'une matrice de distance, et est donc une bonne manière de représenter schématiquement l'éloignement entre deux topics ou entre ceux-ci et leur niveau commun de regroupement. Ainsi, le dendrogramme permet d'exploiter la similarité entre topics afin de construire des classes de topics, de façon à détecter les topics voisins les uns des autres, voire de fusionner certains topics non suffisamment discriminants entre eux (Moretti).

On peut observer sur le dendrogramme deux principaux blocs qui s'agrègent en dernier, constituant les deux classes principales de la bonne partition de notre corpus selon la métrique utilisée : d'une part, une classe que nous appelons « actions publiques », composée de sujets de fond du débat public et des multiples acteurs et institutions qui y participent et d'autre part, une classe qui indique les « conflits directs », notamment tout ce qui a trait aux manifestations.

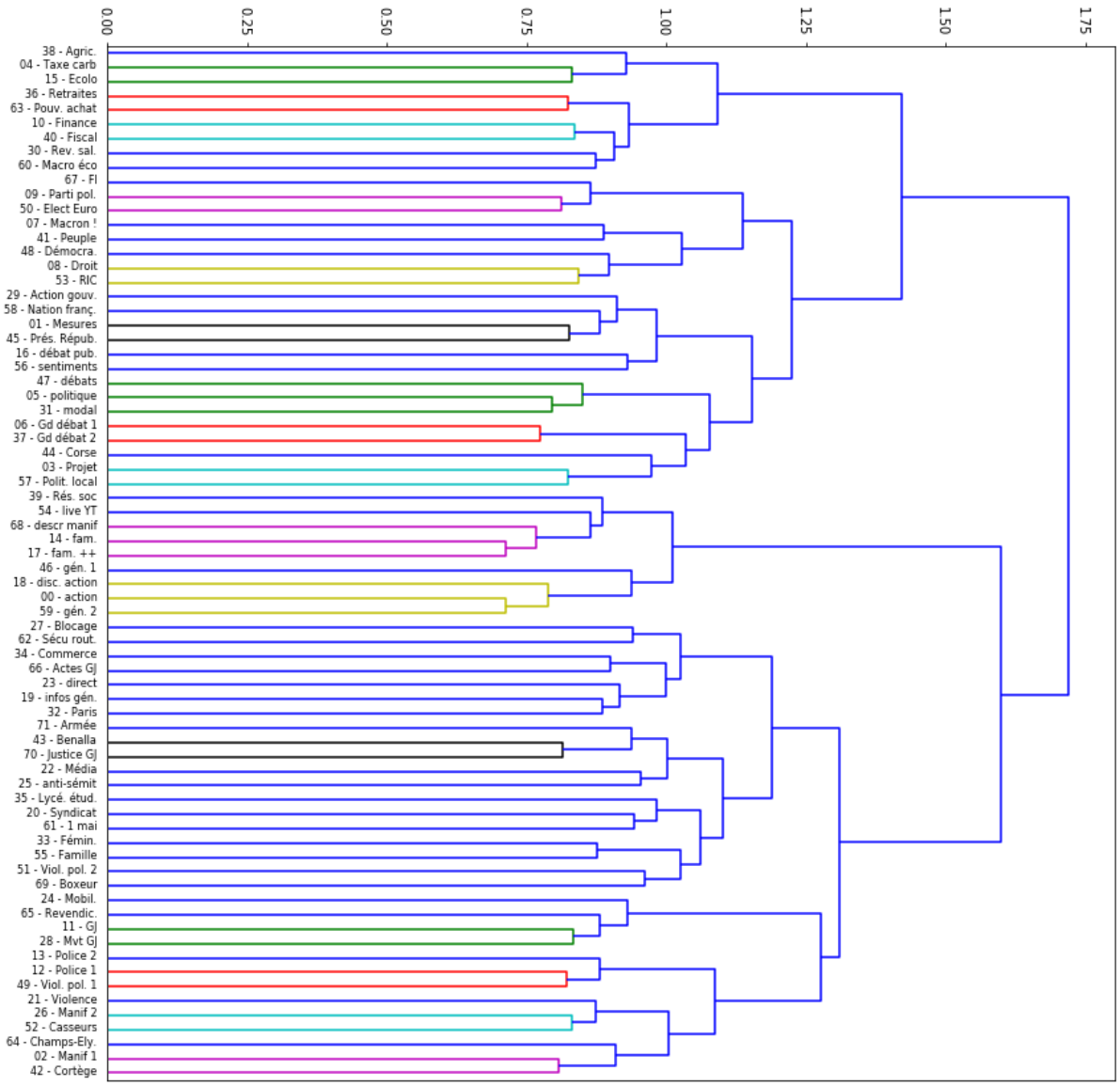

Figure 7 Regroupement hiérarchique ascendant (agglomératif), en utilisant comme distance la valeur 1 - similarité (corrélation de Pearson). La méthode d'agglomération utilisée est celle de Ward, qui permet de mieux identifier des petits regroupements et de ne pas coller arbitrairement entre eux les regroupements. Le seuil de regroupement pour la coloration des branches inférieures du dendrogramme a été pris suffisamment bas pour mettre en évidence la dizaine de groupements les plus pertinents. On lit par exemple que les topics “taxe carburant” et “écologie”, très proches, se sont agrégés dès la première itération, puis ces deux topics s'agrègent dans une deuxième itération avec l'agriculture, etc.

Pour comprendre ces deux principales classes (actions publiques et conflits directs), descendons d’un nœud dans la hiérarchie de la classification, et examinons pour chacune d’elle le tracé de l’arbre en dégageant maintenant les principales classes qui les composent. On observe ainsi que les deux classes que nous venons de présenter sont elles-mêmes décomposées en deux sous-classes. Ainsi l'espace médiatique sur YouTube peut être décomposé en quatre grandes classes de discours :

- ***La classe des topics d’objets de l’action publique.*** On observe d’abord en partant du haut de la liste des topics, une première classe composée uniquement de topics liés à des enjeux sociaux, environnementaux et écologiques : T38 Agriculture, T04 Taxe carburant, T15 Ecologie, T36 Retraites, T63 Pouvoir d’achat, T10 Finance,

T40 Fiscalité, T30 Revenus/Salaire, T60 Macro-Economie. Il s'agit de topics qui renvoient à des topics du débat public.

- ***La classe des topics institutionnels.*** On relève ensuite une classe des topics des institutions qui font et accueillent le débat public : T67 France Insoumise, T09 Parti Politique, T05 Elections Européenne, T07 Macron, T41 Peuple, T48 Démocratie, T08 Droit, T53 RIC, T29 Action Gouvernementale, T58 Nation Française, T01 Mesure, T45 Président de la république, T16 Débat Public, T47 Débat, T05 politique, T06 Grand Débat1, T37 Grand Débat2, T03 Projet, T57 Politique Locale et T44 Corse. Deux topics viennent rompre la cohérence apparente : les topics T31 Modal, renvoyant à des marqueurs de la modalité dans les discussions, et T56 Sentiment, composé d'éléments de qualification de la colère des Gilets jaunes.
- ***La classe des topics de situation d'énonciation.*** En partant du haut de la grande classe des conflits directs, on trouve différents topics indiquant des marqueurs d'oralité : T39 Réseaux sociaux, T54 Live YT, T68 Description Manif, T14 Familier, T17 Fam ++. Ces marqueurs d'oralité renvoient à des situations spécifiques de prise de parole. Comme dans la classe précédente, l'homogénéité de cette classe n'est pas totale. On trouve deux topics très proches et difficilement classables : T00 Action et T18 discours de l'action. Le premier renvoie à des verbes d'action comme "faire", "essayer", "réussir", "décider", "obliger", "pouvoir" (suivi de l'infinitif d'un verbe d'action) qui apparaissent surtout comme des marqueurs d'une attente sociale. Le second indique des marqueurs de l'exhortation à l'action et au changement : "falloir", "électrochoc" et "nouvelle politique" sont les mots et expressions les plus saillants de ce topic. Enfin, c'est aussi dans cette classe que l'on trouve deux topics dont il est impossible de discerner un sens spécifique, ce pourquoi ils sont qualifiés de généraux : T46 General 1 et T59 Général 2.
- ***La classe des topics autour des manifestations.*** Enfin, la classe contenant le plus grand nombre de topics traite des manifestations et des différents éléments qui peuvent leur être associés: T27 Blocage, T62 Sécurité routière, T34 Commerce, T66 Actes GJ, T32 Paris, T43 Benalla, T70 Justice GJ, T22 Média, T25 Anti sémitisme, T35 Lycée étudiants, T20 Syndicat, T61 1er mai, T51 Violence policière 2, T69 Boxeur, T24 Mobilisation, T 65 Revendication, T11 GJ, T28 Mouvement GJ, T13 Police 2, T12 Police 1, T49 Violence policière 1, T21 Violence, T26 Manif 2, T52 Casseurs, T64 Champs Elysées, T02 Manifestation 1, T 42 Cortège. Dans cette classe trois topics surprennent : T33 Féminisme, T55 Famille, et T71 Armée. Nous essaierons plus bas de mieux comprendre cette position. Enfin deux topics de cette classe - T23 Direct et T19 Info générales - correspondent à des topics de situation d'énonciation propres aux manifestations, mais pas seulement. Leur position dans cette classe n'est pas entièrement cohérente.

Le dendrogramme nous permet un premier ordonnancement du corpus en quatre classes qui donnent un aperçu d'ensemble de la topologie de l'espace médiatique sur YouTube, mais quelques topics résistent encore à notre tentative de classement. En fait, les topics *co-existent* d'une manière plus complexe que celle des classes et sous-classes que le dendrogramme a permis de représenter.

Les systèmes thématiques

En effet, en représentant les contenus de notre corpus comme un arbre de divergence de topics, nous avons permis une lecture simple et efficace des topics, mais nous avons trahi la réalité de la composante textuelle des topics : les mots et expressions occupent des positions multiples à la fois dans les documents et les topics. Il nous faut maintenant trouver le moyen de représenter les interconnexions entre les termes et les documents et les termes entre eux. Pour ce faire, la représentation graphique la plus simple que nous pouvons mobiliser est le réseau des topics similaires.

Dans ce cas, les similitudes entre topics peuvent s'observer en considérant leurs co-occurrences au niveau des documents du corpus ainsi que dans le vocabulaire des termes qu'ils mobilisent. Dès lors qu'on affecte à ces similitudes des mesures numériques, on peut également considérer toute combinaison de ces mesures pour tenir compte à la fois des distributions des topics sur les documents et des distributions des mots sur les topics. Une mesure de similarité entre deux topics se calcule en considérant les vecteurs associés aux deux topics, tant sur la matrice documents X topics que sur la matrice topics X termes. Un grand nombre de mesures de similarité entre deux vecteurs sont disponibles parmi lesquelles : le cosinus, le coefficient de corrélation de Pearson, le coefficient de corrélation de Spearman, la divergence de Jensen-Shannon etc. Nous avons utilisé le très classique coefficient de corrélation de Pearson, et ainsi construit une liste de K*(K-1)/2 similarités entre les topics, en considérant à pondération égale tant les similarités au niveau des documents que celles au niveau des termes.

Pour la représentation graphique du réseau des corrélations entre topics, plutôt que d'avoir un graphe illisible (1390 connexions pour 72 topics), nous n'avons conservé que les connexions dont le poids est supérieur à un seuil retenu de façon à obtenir le nombre de connexions minimales pour qu'aucun topic ne soit isolé. On obtient ainsi un réseau plus lisible de 451 connexions, éliminant les structures les moins importantes du réseau de départ.

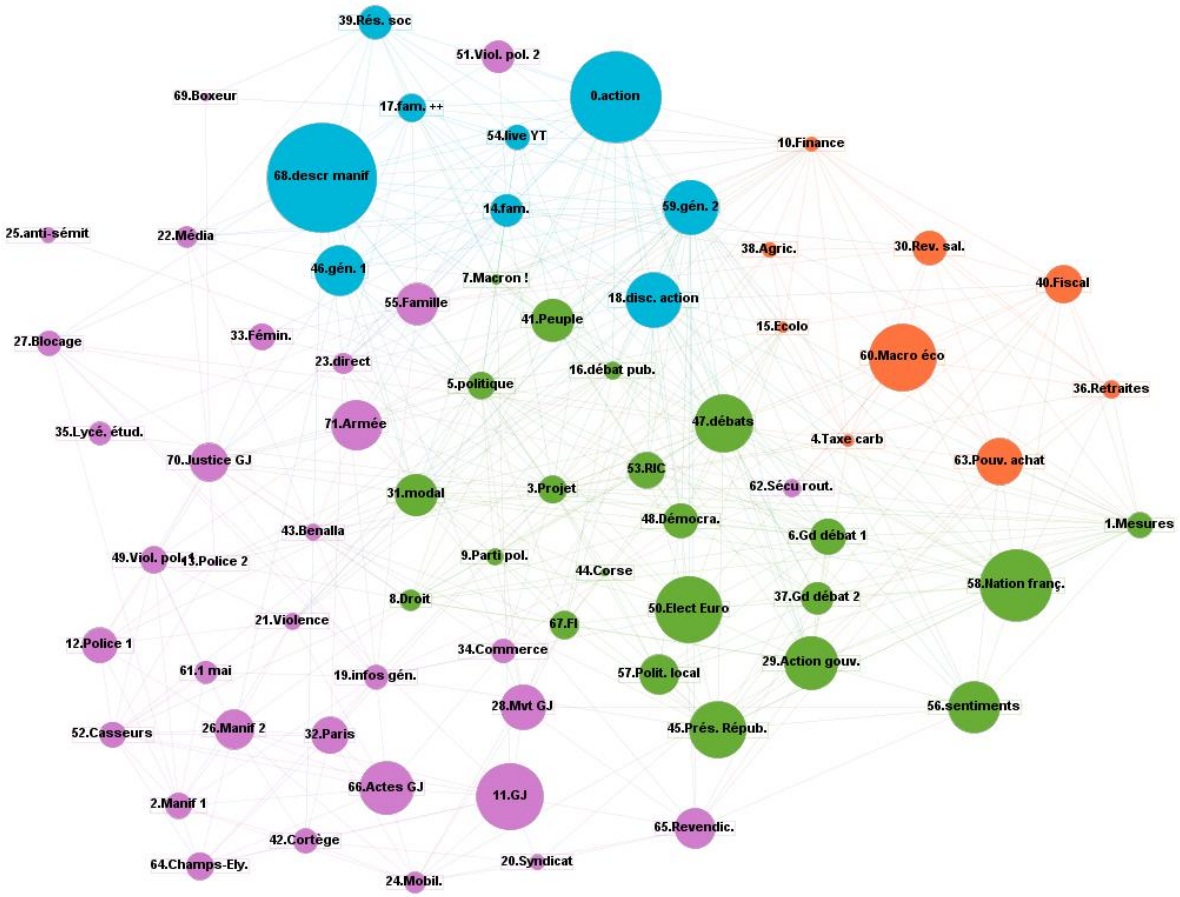

Figure 8 Réseau de topics similaires. Les liens conservés ont un poids > 0,06. Les couleurs correspondent aux classes calculées avec le dendrogramme. En bleu, les topics de situations d'énonciation, en orange les topics des objets de l'action publique, en vert les topics institutionnels et en violet les topics autour des manifestations. L'algorithme de spatialisation utilisé est Force Atlas 2 avec Gephi.

Qu'est-ce que cette visualisation des topics en réseau peut nous apprendre de plus sur le contenu de notre corpus ? Elle apporte une nouvelle synthèse synoptique à partir de laquelle nous pouvons interroger les situations que nous ne parvenions pas à interpréter sur le dendrogramme. Commençons par les trois topics stylistiques qui indiquent des marqueurs d'oralité, mais qui dans le dendrogramme apparaissent dans la classe des topics autour des manifestations (T23 Direct) et celle des topics institutionnels (T56 Sentiment et T31 Modal). Retraçons pour chacun de ces trois nœuds les chemins de connexions vers les autres topics.

Il est intéressant de noter que le T23 Direct départage trois types de situation : le discours présidentiel, le suivi des manifestations à Paris et les chats. Ce positionnement nous montre que le topic « Direct » est un *topic de contexte* qui renvoie à des situations d'oralité très diverses, d'où sa position éloignée de la classe de l'oralité dans le dendrogramme.

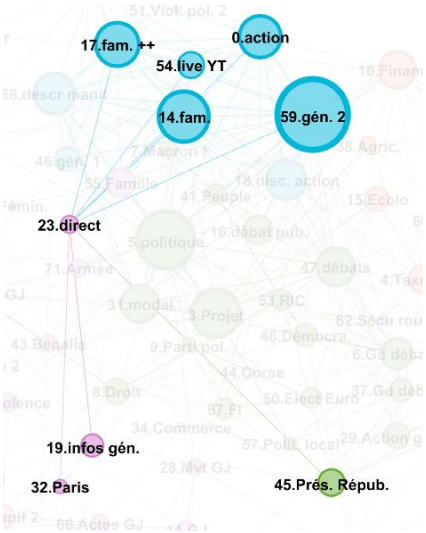

Figure 9 Capture d'écran du voisinage direct du topic T23 Direct.

Le T31 Modal indiquant les marqueurs de modalité aurait aussi toute sa place dans la classe des topics de l'oralité, pourtant le dendrogramme le place dans les topics institutionnels. Ceci n'a rien d'étonnant, mais cette position mérite néanmoins qu'on observe plus en détail la position de ce topic dans le réseau. Nous avons choisi de comparer sa position avec un autre topic qui indique lui aussi des marqueurs de la discussion, à savoir le T54 Live YouTube.

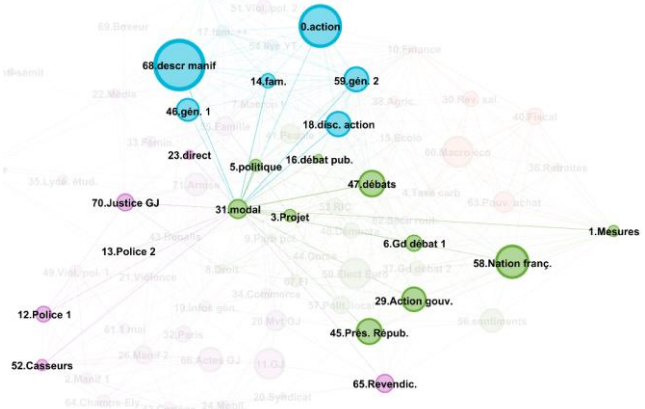

Figure 10 Capture d'écran du voisinage direct du topic T31 Modal.

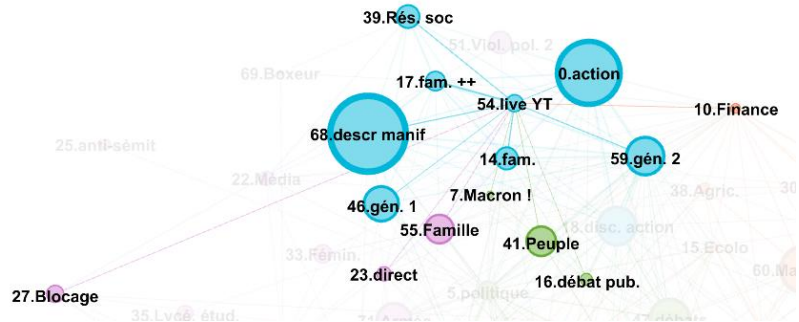

Figure 11 Capture d'écran du voisinage direct du topic T 54 Live YouTube.

Ce qu'il faut retenir de cette comparaison, c'est que ces deux topics produisent des systèmes thématiques très différents : le T31 Modal, plus officiel, se concentre autour des thèmes notoires autour du mouvement (la justice, la police, les casseurs, le président), l'autre plus informel (connexion forte avec le T17 Fam++ qui indique des marqueurs de la vulgarité) se connecte sur la famille et la finance. On trouve ici ce que l'on peut nommer un cadrage médiatique différencié : alors que les discours les plus formels, que l'on peut attribuer à des chaines plus officielles, évoquent les éléments de violence autour du mouvement, les discours les plus familiers sont connectés au sujet de la famille qui renvoie en grande partie à la polémique sur la théorie du genre et à la finance du point de vue des inégalités qu'elle produit.

Le réseau apparaît en effet comme un bon moyen de repérer des cadrages, mais il faut savoir l'interroger. Par exemple, comment repérer les cadrages des raisons de la colère des Gilets jaunes ? Plusieurs topics semblent pouvoir apporter des éléments de réponse à cette question, notamment les T56 Sentiment, T65 Revendication et T29 Action gouvernementale. Dès lors, confrontons les trois réseaux centrés à partir de ces topics, comme on peut le voir dans la figure 12.

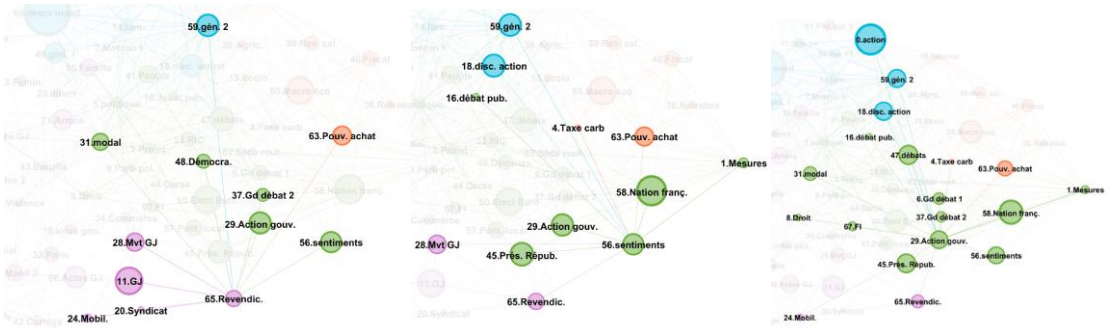

Figure 12 Capture d'écran du voisinage direct, de gauche à droite, des topics T65 Revendication, T56 Sentiment et T29 Action Gouvernementale.

Ce qui est frappant à l'observation de ces graphes, c'est que non seulement le topic « Pouvoir d'Achat » est commun aux trois réseaux égocentrés, mais qu'il est le seul topic parmi les topics objets de l'action publique à être directement lié aux topics de revendication, d'expression de la colère et de l'action gouvernementale. Ainsi le cadrage médiatique autour du problème Gilets jaunes et des solutions à y apporter s'est principalement concentré sur le pouvoir d'achat au détriment d'autre sujets. On peut dès lors se demander quel système thématique forment les autres topics objets de l'action publique, notamment T10 Finance, T60 Macroéconomie, T15 Ecologie et T04 Taxe carbone.

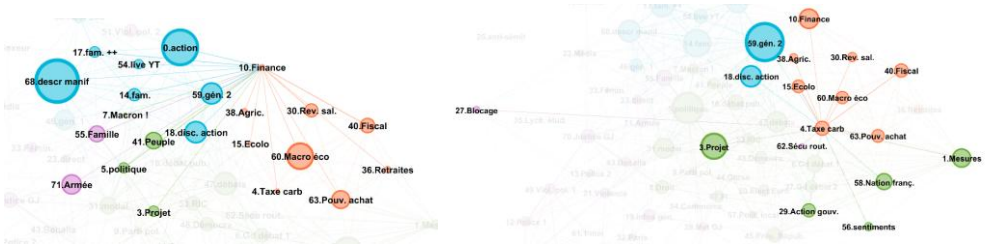

Figure 13 Capture d'écran du voisinage direct du topic T10 Finance, à gauche et T04 Taxe Carburant à droite.

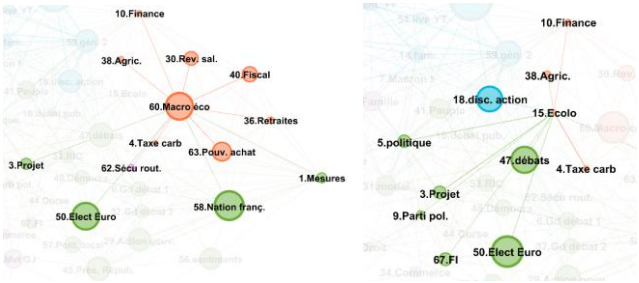

Figure 14 Capture d'écran du voisinage direct des topics T60 Macro-économie, à gauche, et T15 Ecologie, à droite.

La finance et la taxe carbone parviennent à s'interconnecter à l'ensemble des topics objets de l'action publique. En effet, en partant de la finance et de la taxe carbone, on perçoit bien les multiples enjeux interconnectés qui vont du climat au revenu des ménages. En revanche, les topics « Macro-économie » et « Ecologie » forment des systèmes thématiques plus restreints. Plus encore, ils ont seulement en commun la taxe carbone et l'agriculture et ils ne sont pas liés entre eux. Ce qui signifie qu'on parle rarement ensemble de ces deux thèmes qui restent pourtant étroitement connectés dans la réalité. L'abstraction produite par le réseau ne permet pas d'aller plus en détail dans l'interprétation, mais c'est une avancée que de pouvoir montrer cette séparation entre économie et écologie.

Reste un dernier topic qui occupe une position dans le dendrogramme difficile à comprendre : pourquoi l'armée, la famille et le féminisme figurent-ils parmi les topics autour des manifestations, alors qu'ils pourraient avoir toute leur place parmi les topics objets de l'action publique ?

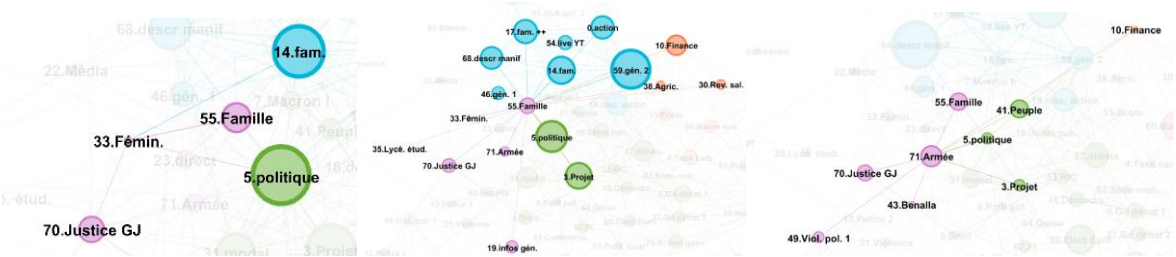

Figure 12 Capture d'écran du voisinage direct, de gauche à droite, des topics T33 Féminisme, T55 Famille et T71 Armée.

Si la connexion du topic de la famille au féminisme est compréhensible, celle qui lie la famille à l'armée est bien moins évidente. Plus encore, la famille s'étend vers la finance, l'agriculture et les revenus des ménages. Pour quelle raison ? La centralité élevée de la famille comme topic est assez inattendue, c'est pourquoi nous avons exploré les titres des vidéos qui ont un score élevé pour ce topic. On remarque que les vidéos de ce topic traitent bien de la famille, mais qu'ils embrassent des sujets très différents allant de la manière dont

les Gilets jaunes font garder leurs enfant les samedis jusqu'à la polémique contre la théorie du genre suscitée par l'activiste Farida Belghoul qui s'est revendiquée Gilets jaunes, en passant par les difficultés financières pour les familles de Gilets jaunes au moment des achats de Noël.
Comme le topic de la famille, le système thématique de l'armée surprend en se déployant sur des topics très différents, les violences policières, la finance, l'affaire Benalla et « Peuple et Pouvoir ». C'est que le topic de l'armée renvoie tout aussi bien aux usages de sentinelles pour la protection des équipements publics qu'aux discussions sur les insurrections à venir, en passant par des discussions conspirationnistes sur le « nouvel ordre mondial ». Le vaste périmètre couvert par ces sujets pourrait expliquer pourquoi il ne s'agit pas d'un topic de l'action publique[4].

Le réseau du féminisme apparaît comme un système plus simple et ses connexions vers la justice, la famille et la politique invite toujours à s'interroger sur sa place dans la classe des topics autour des manifestations. En effet, on observe un système thématique qui semble tourner autour des rapports de genre comme objet de l'action publique (justice, famille et politique). Le réseau ne permet donc pas de comprendre la nature du topic féminisme. Nous apporterons un peu plus bas une réponse à cette question lorsque nous étudierons les évolutions temporelles de ce topic.

Les profils de topics selon les catégories

Le dendrogramme et le réseau sont de bons outils pour comprendre la nature des topics, mais ils ne permettent pas de répondre à la question principale qui motive notre enquête : observe-t-on un agenda médiatique différencié selon les types d'acteur qui occupent l'espace médiatique sur YouTube ? Pour répondre à cette question, il faut chercher les relations existantes entre les topics et les catégories de chaînes.

Nous représentons les relations entre topics et catégories en utilisant un procédé classique : des matrices de corrélation, en particulier des cartes de chaleur sur lesquelles peuvent être appliquées des classifications hiérarchiques selon le sens des corrélations, à la fois pour les entités en ligne et en colonne. La figure 15 montre cette représentation d'ensemble. Le dendrogramme vertical des catégories de chaînes indique une proximité des chaînes selon certains topics. On voit comment les catégories de chaînes forment deux classes distinctes : d'un côté les chaînes qui produisent du contenu de journalistes professionnels opérant au sein de médias traditionnels (catégories Information locale et Médias mainstream) et de l'autre une classe rassemblant les catégories de chaînes YouTube qui sont au cœur de l'élargissement et des reconfigurations de l'espace médiatique - cette classe formant deux sous-classes : Gilets jaune et Contre information d'une part, et Médiations, Médias alternatifs et analyses et Politique, d'autre part.

[4] Notons que cette variété de sujets implique que des termes et expressions liés à ces sujets ont subsisté à la marge et, partant, se retrouveraient dans les relations marginales que le réseau donne à voir. Il faut admettre que les méthodes computationnelles produisent des abstractions à partir desquelles on peut difficilement produire du sens.

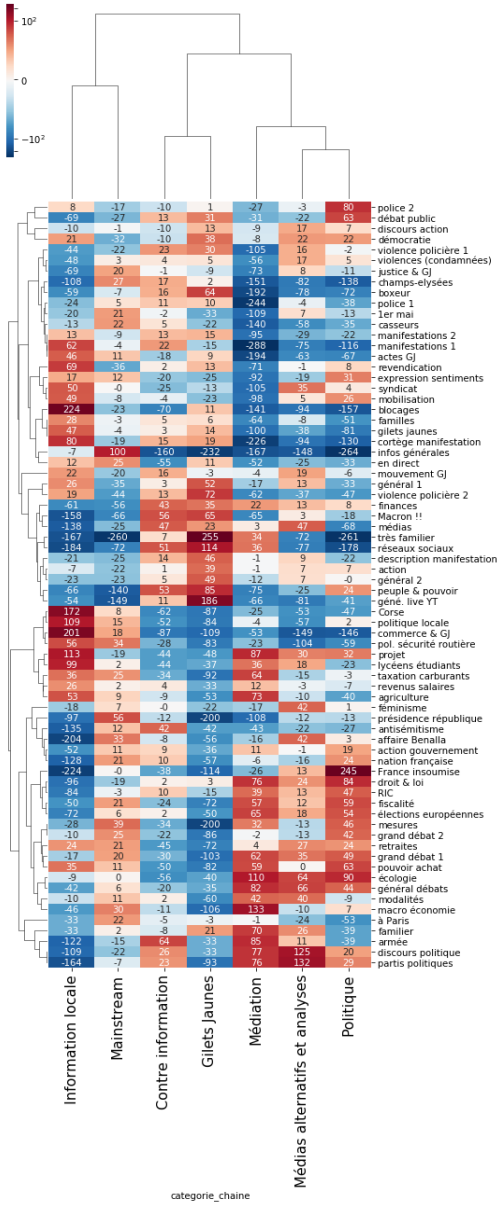

Figure 15 Matrice de corrélation avec classement hiérarchique non supervisé des catégories de chaîne et des topics. Le score indiqué est une mesure d'une déviation odds ratio par rapport à une valeur attendue s'il y avait indépendance entre topics et catégorie de chaîne. Les scores sont visualisés en bleu pour une saillance négative, en rouge pour une saillance positive et en blanc pour un score de saillance nul, donc une absence de corrélation.

Une première lecture de la carte de chaleur peut être faite en listant les topics dont les corrélations positives sont concomitamment élevées pour chacune des trois classes que nous venons de présenter, qui forment, pour rappel, deux binômes et un trinôme de catégories de chaîne. Le binôme des médias locaux et mainstream se constitue par les topics traitant de la Corse, des politiques locales, de l'impact du mouvement sur les petits commerçants et sur la sécurité routière. Pour le deuxième binôme, la proximité des chaînes de Gilets jaunes aux chaînes de contre-information tient surtout aux cinq topics : Boxeur, Live YT, Peuple et pouvoir, Macron et Finances. Enfin le trinôme des chaînes de médias alternatif, médiation et politique s'explique en grande partie par les topics communs suivants : Général, Débat, Ecologie, Grand débat 1, Elections européennes, RIC, Droit et loi. Cette proximité thématique entre catégories de chaîne repose systématiquement sur un petit nombre de topics.

De fait, chaque catégorie de chaîne à un profil de topics qui lui est propre. Faire la comparaison des spécificités thématiques pour chaque catégorie de chaîne est une bonne manière de montrer les cadrages médiatiques différenciés. Les médias locaux ont porté un intérêt particulier à la couverture des manifestations en général, en témoigne la surreprésentation des topics sur les blocages, les cortèges, les commerçants et les revendications. Notons aussi que les violences policières ont été particulièrement abordées à partir des témoignages des Gilets Jaunes (corrélation positive pour Violence policière 2 qui contient beaucoup de marqueurs d'oralité familière et corrélation négative pour Violence policière 1). Il faut souligner que les médias locaux sont aussi ceux qui se concentrent particulièrement sur des topics d'action publique (en particulier l'agriculture), sauf l'écologie qui y est sous-représentée.

Les médias mainstream ont un profil bien particulier : ils se caractérisent par un nombre important de topics aux scores de saillance presque nuls. Ce qui signifie que les médias traditionnels ne se distinguent pas par une sous- ou surreprésentations de topics, hormis pour la sous-représentation des topics de contexte qui renvoie aux styles propres des contenus de réseaux sociaux (souvent très familiers). Ces médias se distinguent néanmoins par une surreprésentation (faible) pour les topics Justice, Champs-Élysées, 1er mai, Casseurs, Direct, Président de la république, Paris. Enfin, dans cette catégorie, le topic Macro-économie est celui qui a le score de saillance le plus élevé parmi les topics de l'action publique, l'écologie restant à un niveau de corrélation nulle.

Les chaînes de contre-information se caractérisent par un non-traitement des topics de fond du débat public, hormis celui de la finance qui atteint le score de saillance maximum pour cette catégorie. Si le dendrogramme sur la matrice indique une similarité de topics entre les chaînes de contre-information et les chaînes de Gilets jaunes, on observe aussi que la contre-information présente des scores de saillance similaires avec les médias mainstream, notamment sur les topics de l'antisémitisme, des casseurs et des Champs-Élysées. Cette similitude tient sans doute à la présence de Russia Today dans la catégorie des chaînes de contre-information qui, de fait, pourrait tout aussi bien appartenir à la catégorie des chaînes mainstream. Cette similitude tient aussi peut-être à la logique systématique de réaction de la contre-information par rapport aux médias mainstream. Notons enfin la spécificité de la contre-information : le score de saillance le plus élevé est celui du topic de l'armée, vient ensuite dans l'ordre décroissant Macron, Peuple et Médias. On retrouve bien ici quelques ingrédients de la rhétorique contre-informationnelle.

La proximité du profil thématique des Gilets jaunes avec celui de la contre-information, telle qu'indiquée par le dendrogramme doit être traitée avec précaution. En effet, si les chaînes de Gilets jaunes partagent avec la contre-information la caractéristique de s'intéresser uniquement à la finance, elles s'en distinguent en accordant un intérêt particulier aux topics Démocratie et Débat public, les Gilets jaunes ayant fait du débat sur la citoyenneté le cœur de leur discussion.

De fait, les chaînes qui atteignent les scores de saillance les plus élevés sur les topics d'objet de l'action publique appartiennent aux catégories médiation, médias alternatifs et analyse, et politique. Parmi ces types de chaînes, la catégorie médiation se distingue néanmoins par les scores les plus saillants pour les topics Macro-économie, Écologie,

Agriculture et Taxation carburant. Soulignons que le topic de l'armée atteint aussi pour cette catégorie de chaînes le score de saillance le plus élevé si on le compare à celui atteint par les autres catégories.

La catégorie de chaînes médias alternatifs et analyses se caractérise par la saillance importante des topics Discours politiques, Partis politiques, Affaire Benalla, Féminisme et Violences condamnées. Notons des corrélations négatives avec les topics Macro-économie, Agriculture, Revenus et salaires et Taxation carburant, et une corrélation nulle avec le topic Pouvoir d'achat. Le topic Gilets jaunes, qui renvoie à des discussions générales sur le sens à donner au mouvement, y est largement sous-représenté. Dès lors, tout porte à croire que ce qui a intéressé surtout les médias alternatifs durant la première année du mouvement relève surtout du scandale, de la dénonciation et des luttes partisanes.

Enfin, les chaînes politiques se distinguent par une surreprésentation de la France Insoumise qui tient à la surproduction de contenus du parti de Jean Luc Mélenchon et de ses députés. Les partis politique sur YouTube ont concentré leur ligne éditoriale sur les topics Pouvoir d'achat, Grand Débat, Expression sentiments, Débat public et Police 2 (moyens donnés aux policiers). Ce que révèle ainsi les topics, c'est la stratégie de positionnement des partis, notamment de la France Insoumise, en porte-parole des Gilets jaunes, mais sans attachement spécifique aux définitions des objets de fond du débat public.

Les profils de topics selon les actes du mouvement

Une dernière analyse s'impose, celle de l'évolution temporelle des topics. Observe-t-on des topics caractéristiques de certaines périodes ? Comment les topics se distribuent-ils dans le temps ? Pour répondre à ces questions, nous avons mobilisé une analyse de contingence afin de calculer les corrélations entre les topics et les actes de GJ. La méthode est la même que celle mobilisée pour les catégories de chaîne, mais nous avons représenté la distribution des topics avec des histogrammes de scores de saillance (déviation *odds ratio*), plutôt qu'avec des matrices. On produit ainsi une visualisation de la façon dont les topics sont répartis sur les vidéos selon les actes, et inversement. Cette méthode nous permet de mesurer l'intensité avec laquelle un topic est traité selon les actes. Les visualisations suivantes comparent l'évolution des scores de saillance pour quatre topics à la fois. Nous y avons systématiquement ajouté un graphe d'évolution des volumes associés à ces topics, ce qui apporte une information complémentaire, la distribution des volumes de vidéos selon les topics dans le temps.

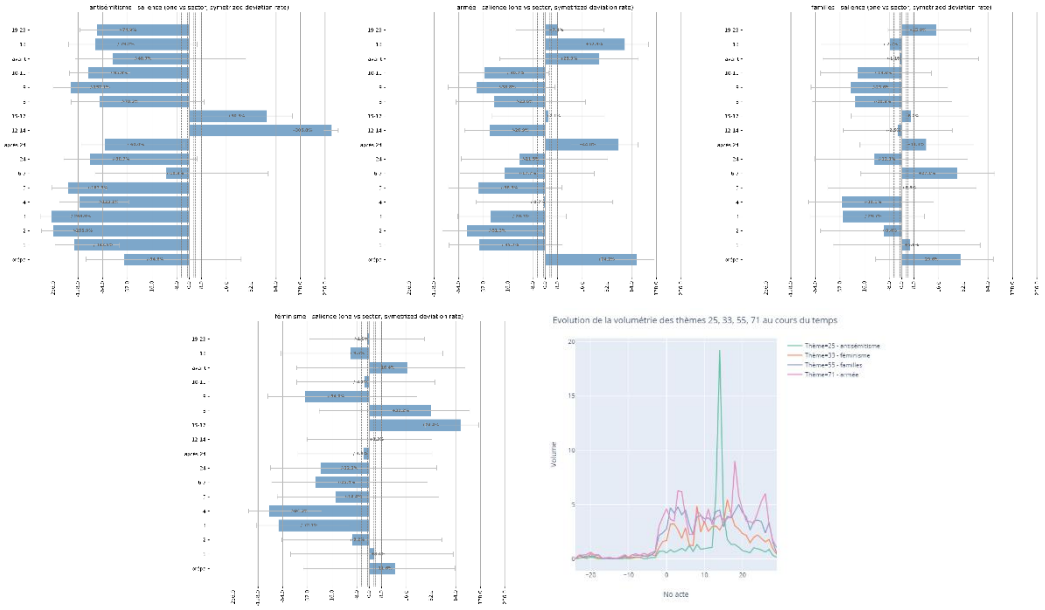

Figure 16 Profil temporel selon les actes pour les topics (de haut en bas) Antisémitisme, Armée, Famille et Féminisme. La saillance du topic par acte est indiquée par un score de déviation symétrique négatif ou positif. Attention, les actes ne sont pas ordonnés dans le sens du temps sur les profils. En bas à droite, l'évolution de la volumétrie des quatre topics.

Commençons par les topics Antisémitismes, Armée, Féminisme et Famille. Le topic Antisémitisme se concentre autour de l'acte 15, suite aux injures antisémites dont a été victime le philosophe Alain Finkielkraut lors d'une manifestation. C'est donc un sujet épisodique, et non pas un thème structurant du mouvement. En effet, le volume de ce topic est quasi nul en dehors de cette période et atteint un score de saillance important et une pointe exceptionnelle lors de cet évènement (c'est la pointe de loin la plus haute par rapport aux autres topics). Le topic de l'Armée quant à lui est plus constant dans le temps, avec une légère surreprésentation au moment de l'acte 18 qui s'explique par l'annonce gouvernementale de la mobilisation des militaires pour protéger les bâtiments officiels lors des manifestations. Le topic du Féminisme présente un profil à plusieurs saillances. Lors du 1er acte, le topic est surreprésenté en raison du recouvrement des Gilets Jaunes avec le mouvement « NousToutes ». Le thème réapparaît à partir de l'acte 8 suite à une manifestation dominicale de femmes Gilets jaunes. La pointe autour de l'acte 15 correspond aux discussions qui ont suivi la journée de la femme du 8 mars. Sur la période étudiée, l'évolution du volume de ce thème est plus ou moins croissante jusqu'au 8 mars, puis diminue avec l'essoufflement progressif des manifestations. On comprend mieux pourquoi le féminisme est apparu dans la classe des topics des manifestations du dendrogramme, et non pas dans celle des objets de l'action publique.

Les topics que nous avons qualifiés d'objet de l'action publique, notamment l'Ecologie, la Macro-économie, la Finance et l'Agriculture, ont-ils des profils temporels moins épisodiques que ceux que nous venons de commenter ? Commençons par la macro-économie qui apparaît en surreprésentation en amont, puis après l'acte 24. Cette surreprésentation ne signifie pas pour autant une absence de ce topic en dehors de ces périodes. En effet, nous observons que le topic de la macro-économie est un topic représentant un volume de vidéos important et constant dès l'acte 1. La Macro-économie est donc un topic quasi permanent dans l'espace public autour du mouvement. En revanche, le topic de l'écologie est

surreprésenté lors des actes 2, 3 et 4 puis à nouveau sur la période allant de l'acte 15 à l'acte 18. Il reste néanmoins à un niveau relativement faible en dehors de ces périodes. Notons que la période allant de l'acte 15 à l'acte 17, qui correspond aux dernières semaines avant la clôture du Grand Débat, est caractérisée par une saillance forte des trois topics, Ecologie, Agriculture et Finance. Cette saillance s'explique donc par l'intensification des débats de fond due au Grand Débat.

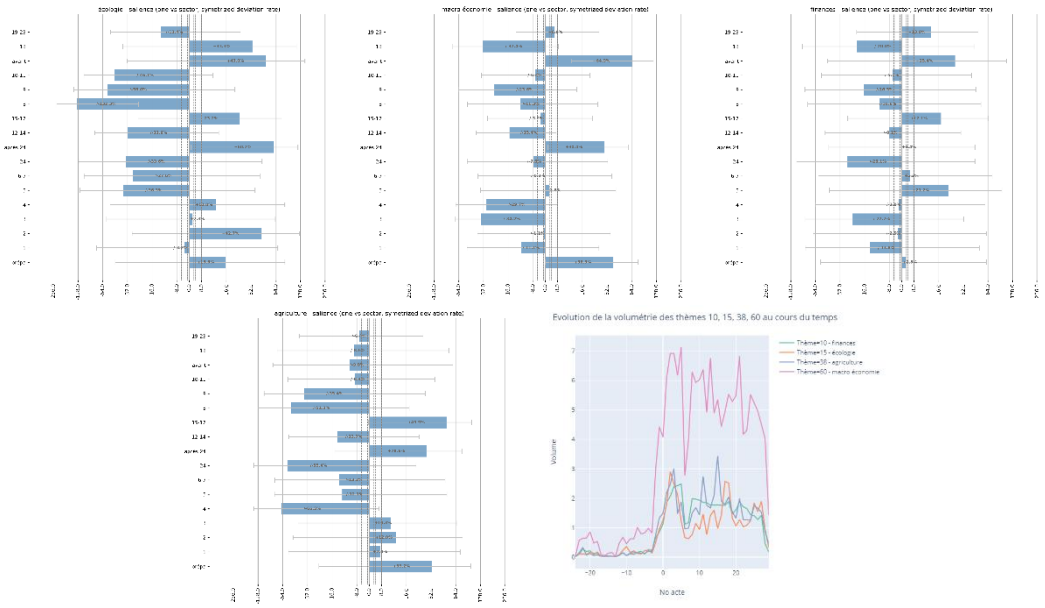

Figure 17 Profil temporel selon les actes pour les topics (de haut en bas) Ecologie, Macro-économie, Finance, Agriculture. La saillance du topic par acte est indiquée par un score de déviation symétrique négatif ou positif. Attention, les actes ne sont pas ordonnés dans le sens du temps sur les profils. En bas à droite, évolution de la volumétrie des quatre topics.

Passons aux quatre topics objet de l'action publique qui ont été aux origines du mouvement : la taxation des carburants, le pouvoir d'achat, les revenus et, dans une moindre mesure, les retraites.

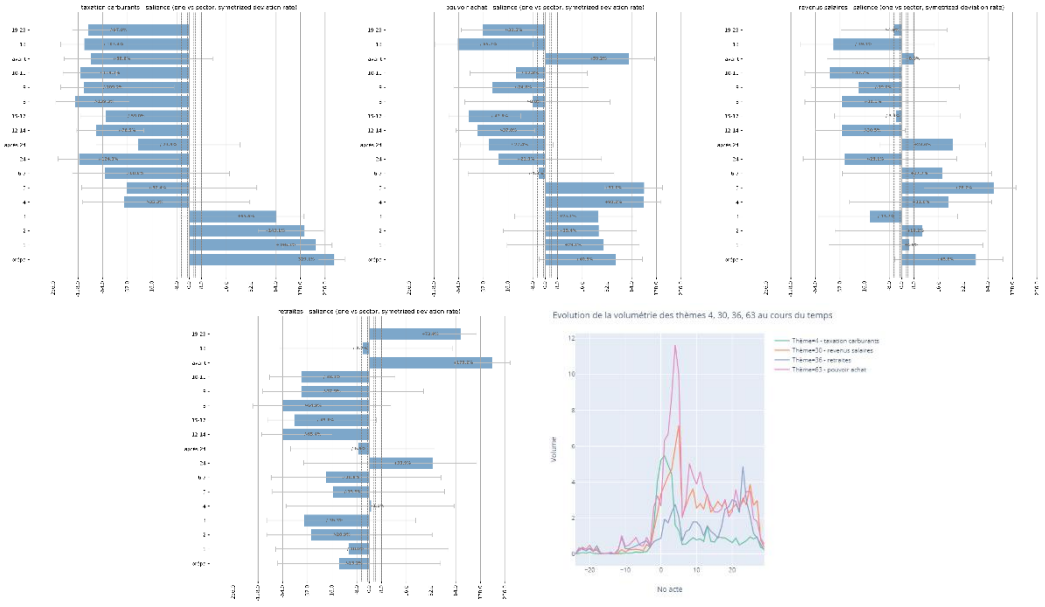

Figure 18 Profil temporel selon les actes pour les topics (de haut en bas) Taxation carburants, Pouvoir d'achat, Revenus et salaire et Retraites. La saillance du topic par acte est indiquée par un score de déviation symétrique négatif ou positif. Attention, les actes ne sont pas ordonnés dans le sens du temps sur les profils. En bas à droite, évolution de la volumétrie des quatre topics.

La taxation du carburant a cessé d'être caractéristique des débats sur les Gilets jaunes dès le troisième acte, le pouvoir d'achat à partir du cinquième et les revenus à partir du septième. Notons que sur le plan volumétrique, ces trois topics restent néanmoins beaucoup plus importants que le topic de l'écologie sur les périodes avec une saillance négative. Enfin, si le topic des retraites est particulièrement surreprésenté autour de l'acte 19, une phase postérieure à la clôture du grand débat, son volume est quasi équivalent à celui de l'écologie en dehors de cette période de saillance.

Pour terminer notre exploration, observons les topics institutionnels, en particulier, Peuple et pouvoir, Démocratie, Débat public et Grand débat 2. Ce qui est le plus manifeste dans leurs profils temporels, c'est la saillance des topics décalée d'un acte : d'abord Peuple et pouvoir qui est le topic le plus contestataire, et qui émerge et disparait progressivement entre l'acte 2 et l'acte 10 ; ensuite vient le topic Démocratie, d'une nature conceptuelle plus élevé, qui démarre autour de l'acte 5 et s'arrête aussi à l'acte 10 ; puis le topic Débat public, qui décrit le déroulé des consultations d'Emmanuel Macron, débute autour des actes 6 et 7 et s'estompe à l'acte 18 ; et enfin le topic Grand Débat, à partir de l'acte 8 jusqu'à l'acte 23, qui renvoie à des éléments plus techniques de fonctionnement de la plateforme participative et de contenu du débat.

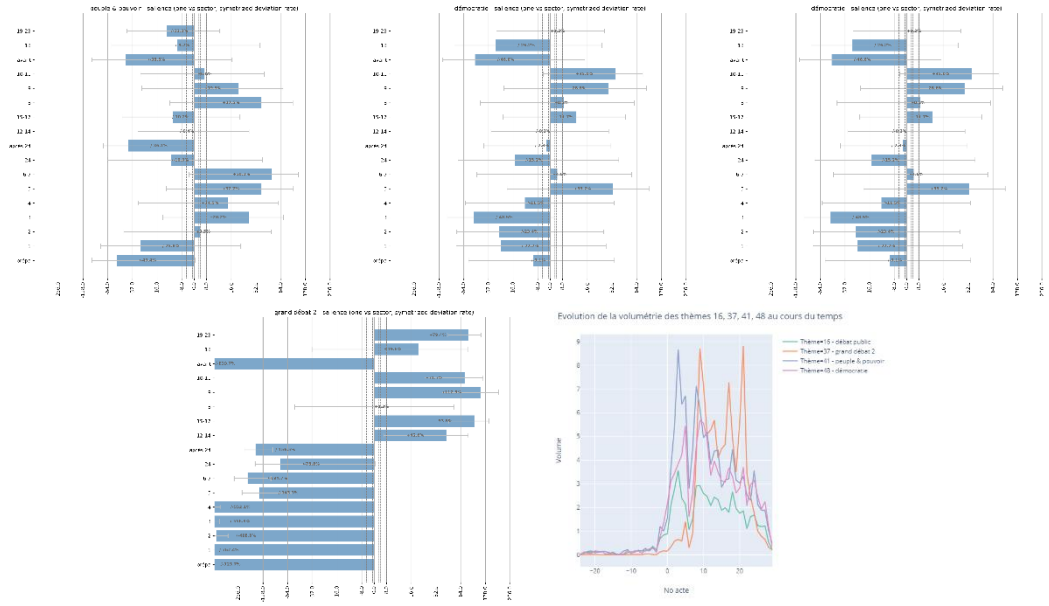

Figure 19 Profil temporel selon les actes pour les topics (de haut en bas) Pouvoir et peuple, Démocratie, Débat public et Grand débat. La saillance d'un topic par acte est indiquée par un score de déviation symétrique négatif ou positif. Attention, les actes ne sont pas ordonnés dans le sens du temps sur les profils. En bas à droite, évolution de la volumétrie des quatre topics.

Si ces quatre topics traitent de questions de citoyenneté, ils ont chacun un statut différent : alors que Peuple et pouvoir ainsi que Démocratie correspondent à des discussions critiques, voire conceptuelles, autour du système politique et de la démocratie représentative, les topics Débat public et Grand débat 2 portent sur le déroulement de l'initiative participative mise en place par le gouvernement. On observe ainsi comment le « Grand Débat » a eu un

effet progressivement annihilateur sur les topics critiques du système politique et d'une manière plus générale, un effet de cadrage sur l'espace médiatique des Gilets jaunes.

Discussion

Telles sont donc les informations que le dendrogramme, le réseau, la matrice de corrélation et les profils temporels nous ont données à voir. Pour intéressant que soient ces résultats, ont-ils apporté des connaissances nouvelles sur le traitement médiatique des Gilets jaunes ? La réponse est évidemment mitigée. Les 72 topics, et les quatre manières de les représenter, ont pu corroborer ce que les spécialistes des médias ont déjà montré sur le traitement médiatique des mouvements sociaux en général, à savoir une surreprésentation de la violence contestataire et spectaculaire dans les médias traditionnels ; des médias alternatifs proposant des informations différentes de celles des médias dominants, mais dépendantes de l'agenda militant. Jusque-là, rien de nouveau dans notre étude. Mais de même que l'analyse par les topics model corrobore par certains aspects les études existantes sur les rapports entre médias, plateformes numériques et mouvements sociaux, on peut dire que réciproquement ces recherches prouvent la fiabilité des topic models et la pertinence de partir de YouTube comme terrain d'enquête. L'un des apports de cet article est d'avoir montré que les sous-titres des vidéos de YouTube forment un matériau sur lequel on peut faire enquête pour analyser l'espace médiatique. Notre étude parvient en effet à mettre en évidence l'espace médiatique atour du mouvement selon quatre classes de topics : les objets de l'action publique, les institutions, les manifestations et les marqueurs d'oralité. Comprendre la cohérence de ces quatre classes aura permis de réduire la complexité d'une analyse en 72 topics d'un traitement médiatique.

S'il y a une nouveauté dans cet article au regard des travaux existants sur le traitement médiatique des Gilets jaunes (Souillard et al., 2020), elle se situe au niveau des topics des objets de l'action publique et des topics institutionnels. Pour les topics des objets de l'action publique, on peut retenir trois résultats importants : ces topics sont caractéristiques des chaînes de médiation, ce qui signifie que la vulgarisation politique a joué un rôle clef dans le traitement médiatique du mouvement en y apportant des éléments de fond ; l'écologie et la macro-économie sont déconnectés et l'écologie est restée marginale relativement à la macro-économie ; et si les thèmes fondateurs du mouvement (pouvoir d'achat, taxation des carburants, revenus et salaires, et retraite) ont diminué progressivement en intensité, ils occupent un volume largement supérieur à celui de l'écologie tout le long de la période étudiée. Au sujet des topics institutionnels, il nous semble particulièrement important de retenir les topics autour des rapports des citoyens à leur système politique. Les Gilets jaunes se sont en grande partie mobilisés autour de revendications sur la démocratie. Ces revendications, visibles dans notre corpus, ont aussi évolué au fil des actes. Nous l'avons montré, le Grand Débat semble avoir eu un effet important, dans l'espace médiatique, sur la structuration des discours autour de la critique de la démocratie représentative. Autrement-dit, tout porte à croire que l'Etat soit parvenu avec ce dispositif à recadrer l'espace

médiatique. C'est peut-être même le principal effet politique du grand débat – « si vous voulez avoir une prise sur l'espace médiatique, créez une plateforme participative[5] ».

Enfin, nous voudrions souligner nos résultats autour des chaines de contre-information dans un contexte où les médias traditionnels ont mis en avant des accointances douteuses de Gilets jaunes pour les médias conspirationnistes et xénophobes (voir par exemple les dénonciations de (Bornstein, 2019)). Nous montrons que si les profils thématiques des chaînes Gilets jaunes et de contre-information ont quelques points communs, cela ne porte pas sur des topics conspirationnistes pour autant. Ces deux catégories de chaîne présentent aussi des divergences importantes, notamment l'intérêt spécifique des Gilets jaunes pour les sujets de citoyenneté que n'ont pas les chaînes de contre-information. La proximité thématique observée des chaînes de contre-information et des médias mainstream invite aussi à des explorations plus approfondies sur les rapports d'interdépendance de ces deux catégories de chaîne dans les situations de crise.

Conclusion

Ainsi, cet article ouvre une perspective novatrice : explorer YouTube à partir d'une analyse non supervisée peut être un moyen puissant de totalisation statistique pour faire émerger un point de vue général sur l'espace médiatique. Mais si la production de contenus médiatiques, autour d'un mouvement social notamment, est un enjeu de quantification, elle est rarement appréhendée comme un objet de statistique. Pourtant, l'analyse de la production médiatique procède d'une forme de statactivisme (Bruno and Didier, 2014). En effet, les méthodes de l'analyse quantitative de contenu qui classent et comptent la production médiatique ont été mobilisées durant le mouvement des Gilets jaunes comme pratiques statistiques pour critiquer et s'émanciper de la réalité médiatique construite par les médias.

Il existe donc deux réalités médiatiques[6] : celle produite par les médias en enquêtant sur le monde social (réalité médiatique n°1) et celle produite par l'analyse de contenu qui en mesurant la production des médias fige une « représentation critique du traitement médiatique » (réalité médiatique n°2).

C'est cette deuxième réalité médiatique que nous avons construite dans cet article. De ce point de vue, l'analyse quantitative de contenu médiatique par les topics models peut être discutée du point de la sociologie de la quantification : l'analyse des topics revient à mettre en place une nouvelle construction via le déploiement d'une infrastructure de classification des contenus. Si la réalité médiatique (n°2) n'est pas accessible directement, mais par l'intermédiaire d'algorithmes, il devient urgent de faire de l'analyse quantitative de contenu un instrument officiel - une statistique publique - systématique et harmonisée pour faire de l'espace médiatique un objet de commune mesure. YouTube et les topics models sont de

[5] On s'inspire ici de la célèbre formule de Clémenceau, « Si vous voulez enterrer un problème, nommez une commission ». La plateforme participative se substituerait-elle à la forme « commission » qui incarne le mieux l'espace politico-administratif de la démocratie représentative ?

6 En s'inspirant de Boltanski, on distingue ici le "monde médiatique", que l'on peut saisir, et la "réalité médiatique", construite par l'analyste qui se donne pour mission d'explorer le traitement médiatique (Boltanski, 2009).

bons candidats pour construire cet espace d'équivalence et le langage commun permettant de débattre des agendas médiatiques et de leurs effets sur l'espace public.

Bibliographie

Arora S, Ge R, Moitra A, 2012. Learning topic models – going beyond SVD. In FOCS (pp. 1–10), IEEE Computer Society

Blei D, 2012. Probabilistic Topic Models, communications of the ACM, vol. 55, no.4

Blei D, Lafferty J, 2006. Dynamic topic models. In Proceedings of the 23rd International Conference on Machine Learning, pages 113–120

Blei D, Lafferty J, 2007. A correlated topic model of Science. Annals of Applied Statistics, 1(1):17–35

Blei D, Ng A, Jordan M, 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022

Boltanski L (2009) *De la critique: précis de sociologie de l'émancipation*. Gallimard.

Bornstein R (2019) En immersion numérique avec les « gilets jaunes ». *Le Debat* n° 204(2). Gallimard: 38–51.

Bruno I and Didier E (2014) *Statactivisme*. Paris: Zones.

Cardon D and Granjon F (2010) *Médiactivistes*. Presses de Sciences Po. Available at: https://www.cairn.info/mediactivistes--9782724611687.htm (accessed 24 July 2020).

Cointet J-P and Parasie S (2018) Ce que le big data fait à l'analyse sociologique des textes. *Revue francaise de sociologie* Vol. 59(3). Presses de Sciences Po: 533–557.

Deerwester S C, Dumais S T, Landauer T K, Furnas G W, Harshman R A, 1990. Indexing by latent semantic analysis, Journal of the American Society of Information Science, 41(6), 391–407

Dieng A B, Ruiz F J R, Blei D, 2019. Topic Modeling in Embedding Spaces. In arXiv:1907.04907

DiMaggio P, Nag M, Blei D, 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding, Poetics. 41. 570–606

Evans J A, Aceves P, 2016. Machine Translation: Mining Text for Social Theory, Annual Review of Sociology, Vol. 42:21-50

Ferron B (2019) Mouvements sociaux : le jeu médiatique en vaut-il la chandelle ? Available at: http://theconversation.com/mouvements-sociaux-le-jeu-mediatique-en-vaut-il-la-chandelle-128139 (accessed 24 July 2020).

Granjon F (2014a) *62. Médias dominants, mouvements sociaux et mobilisations informationnelles*. *Histoire des mouvements sociaux en France*. La Découverte. Available at: https://www.cairn.info/histoire-des-mouvements-sociaux-en-france--9782707169853-page-681.htm (accessed 24 July 2020).

Granjon F (2014b) Citoyenneté, médias et TIC. *Reseaux* n° 184-185(2). La Découverte: 95–124.

Granjon F (2018) Mouvements sociaux, espaces publics et usages d'internet. *Pouvoirs* N° 164(1). Le Seuil: 31–47.

Greene D, Cross J P, 2017. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. Political Analysis, 25(1), pp 77-94

Greene D, O'Callaghan D, Cunningham P, 2014. How Many Topics? Stability Analysis for Topic Models, ECMLPKDD'14: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Volume Part I, pp 498–513

Hofmann T, 1999.Probabilistic latent semantic analysis. In Proc. 15th Conference on Uncertainty in Artificial Intelligence, pp. 289-296

Ludovic, L., & André, S. (1994). Statistique textuelle. Paris, Dunod.

Lee D D, Seung H S, 1999. Learning the parts of objects by non-negative matrix factorization, Nature, 401, 788–791

Lindstedt N C, 2019. Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017, Social Currents, Volume 6 issue 4, pp. 307-318

McCombs ME and Shaw DL (1972) THE AGENDA-SETTING FUNCTION OF MASS MEDIA. *Public Opinion Quarterly* 36(2). Oxford Academic: 176–187. DOI: 10.1086/267990.

Mimno D, Wallach H M, Talley E, Leenders M, McCallum A, 2011. Optimizing semantic coherence in topic models. In Proc. of the Conf. on empirical Methods in Natural Language Processing, pages 262-272

Moretti F (2016) *La Littérature au laboratoire*. 1st ed. Les Éditions d'Ithaque.

Pinto S, Albanese F, Dorso C O, Balenzuela P, 2019. Quantifying time-dependent Media Agenda and public opinion by topic modeling, Physica A: Statistical Mechanics and its Applications, Volume 524, Pages 614-624

Poels G and Lefort V (2019) « Gilets jaunes » : une médiatisation d'une ampleur inédite. *La Revue des Médias* (13). Available at: http://larevuedesmedias.ina.fr/gilets-jaunes-mediatisation-chaines-info-twitter (accessed 24 July 2020).
Roberts M, Stewart B, Tingley D, 2014. stm: R package for structural topic models. R package version 1.1.3.

Röder M, Both A, Hinneburg A, 2015. Exploring the space of topic coherence measures. In Proceedings of the eighth ACM international conference on web search and data mining, WSDM '15 (pp. 399–408)

Sebbah B, Loubère L, Souillard N, et al. (2018) *Les Gilets jaunes se font une place dans les médias et l'agenda politique*. Research Report. Laboratoire d'Etudes et de Recherches Appliquées en Sciences Sociales. Available at: https://hal-amu.archives-ouvertes.fr/hal-02120478 (accessed 24 July 2020).

Sebbah B, Loubère L, Souillard, N, et al. (2019) *La dilution des Gilets jaunes dans l'agenda médiatique et politique*. Rapport de recherche. Laboratoire d'Etudes et de Recherches Appliquées en Sciences Sociales. Available at: https://www.histoiredesmedias.com/Etude-La-dilution-des-Gilets.html (accessed 24 July 2020).

Tsur O, Calacci D, Lazer D. 2015. A frame of mind: using statistical models for detection of framing and agenda setting campaigns. Proc. 53rd Annu. Meet. Assoc. Comput. Linguist. 7th Jt. Conf. Nat. Lang. Process. Int., Beijing, July 26–31, pp. 1629–38

Wesslen R, 2018. Computer-Assisted Text Analysis for Social Science: Topic Models and Beyond, ArXiv abs/1803.11045

Annexe 1

Tableau listant les chaines du corpus par catégorie. Entre parenthèse, le nombre de vidéos par chaîne.

Contre-information
RT France (1726), Fred Bes (236), Marie Louise Peigné Richard (198), Chaîne officielle- TVLibertés (179), L'informateur (143), TANRESI CENSUREEparDesFDP (113), Ça Zap ! (95), Ça crève les yeux (94), Polémique & Blah Blah (72), CitoiCitoyen (69), Eric Perroud 63 (68), Le Frexit (67), Agence LDC News (65), Christophe Cros Houplon (64), Riposte Laïque 2 (64), Le Stu-Dio (63), le veilleur non silencieux (61), VU FranceTV (61), NOP (60), street 45 (58), TV Patriotes (58), Hyper Crazy 9 (54), REAL POLITIC (53), Melis Ann (52), jayslem jayslem (49), Saber Solo Radio Online (49), philippe jandrok (45), Nouvelles (44), CDjamelito (43), Morgan Priest (40), Salim Laïbi (40), Le Libre Penseur (40), Jasper Mader (40), Campagnol tvl (39), Dieudonné Officiel (38), Boulevard Voltaire (38), Le Cercle Richelieu (36), Abu-Ayyub Cédric-Ali (36), J'suis pas content TV (34), ERTV Officiel (34), Retransmission (33), Jean-Yves Le Gallou (32), JC2R Officiel 2 (31), Margot Velazquez (31), DarnaTelevision (30), jayslem officiel (29), Frederic Delavier (29), Droitards Méchants (29), Michel Drac (27), Riposte Laïque (27), La Mite dans la Caverne (26), TEPA (26), ADBK TV (26), BricoLePatriote (26), l'Actu (26), Exister en Liberté (23), Patrick LAROVERRE (23), LIB TROPIQUES (22), L'Esprit franc - Virginia Vota (22), L'Aile à Stick (22), SAINT LYS France (22), Stop Mensonges (21), Investig'Action (21), Cercle Aristote (21), LA VOIX DU PEUPLE (21), Planetes360 (20), radio 17/11 (20), Farida Belghoul (20), PAS LINFO (20), jim le reveilleur (19), La Voix de Portici (18), Bernard Bordas (17), Antyss77 (17), ZAP DU BUZZ (17), Grand Angle (16), PressTV Français (16), Culture Populaire (16), wanted info (15), Adrien Gaugin (15), Le Pixel Mort (15), Commandant AUBENAS (15), Papi Chriis (14), Mika team pro (14), Arti Bus (14), Médias-Presse-Infos (14), AGORA (13), did B (13), jayslem jayslem 2 (13), Le Retour aux Sources Éditeur (13), CHEIKH DIENG OFFICIAL (13), Stratpol (13), Georgia SwitchnewsTV (13), Jean-Paul Miniscloux (13), Assimil TV (12), Joharno (12), Balance news (12), Action Française (12), France - FR (12), Brefnews info (12), forest f (11), lachainedevv (11), Tatiana Ventôse (10), BTLV Le média complémentaire (10), Le Fil d'Actu - Officiel (10), bordcham (10), Artémisia Collège (9), piero san giorgio (9), Vive L'Europe (9), Batdaf 1 (9), GUDRUN Z (8), Ac Floraison (8), jean bricmont (7), Monde Français (7), Délivrance (7), NURÉA TV - Le média qui met les Mystères en Lumière (7), Emmanuel et Sandrine Survivalistes (7), Valéry Coquant (6), Factuelle 66 (6), Radiopariman2 (6), jcrabiller (6), Komrad (6), parole Libre. (6), Libre Opinion (6), Nicolas MJ (6), Actu Médias (5), ANONYMOUS ALTRIA (5), CIA Chaine info d 'une Alien (5), Survivaliste Bushcrafter (5), Reconquête Française X (5), Terra Bellum (5), Michel Taupin, résistant (5), Alpha 77 (5), Stéphane Blet (5), Sortez de la Matrice, la chaîne de Vahine (4), Stéphane Edouard (4), Marc Herstalle (4), lesnonalignes (4), CONTRE-PROPAGANDE (4), Dissident Officiel (4), Thebengale96 (4), Aciiderixx (4), Joël Nadon (4), koi de 9 (4), GRIN2KAF PROD (4), Guy Fawkes (4), Forum France (4), L'Opinion éclairée (3), JE M'INFORME TV (3), En Mode REPLAY (3), SofZilog (3), Laurent Martinez (3), Complotiste Kawaii (3), KILLUMINATY SMG OFFICIEL (3), Clash Politique (3), noisiveleT (3), HelloHelloInfo (3), Chimical Spray (3), Tukabel Tukabel (3), Aldo Sterone (3), Dajjal Magazine (3), SACR TV (3), Institut Iliade (3), lesurvivaliste (2), Nouvelles VN.7 (2), jmichel2you (2), Pagans TV (2), Viking Eco (2), Pierre Jovanovic - La Revue de Presse (2), Peter Moore (2), RASTA PRESIDENT (2), Alexandre Lebreton (2), Kontre Kulture (2), Christian Cotten (2), Fr_News - Toute l'info en vidéo (2), avada_kedavra (2), instinCroma (2), bah voyons! (2), BioticTV (1), ERTV Rhône-Alpes (1), Bourquin Jean-Charles (1), Historique TV (1), Nico & Mariana (1), CHL.TV (1), Paul disciple du Christ (1), Islamotion (1), RING (1), L'HEURE DE SE REVEILLER (1), Nouvelles 24h (1), FalloutVideo & Actus (1), blondewoman1 (1), wat ari (1), Abel Chemoul (1), Teddyboy RSA (1), Vue Autrement (1), Pavalek Norajovitch (1), VérítésQuiDérangent (1), Investig Info (1), code- Rno (1), ER LILLE (1), Télévision Madame l'Afrique (1), La Matrice (1), ??cent soixante-quinze?? (1), reinformation.tv (1), bord eaux (1), Studio Pracoz (1), Génération Identitaire (1), La France En Colère (1), Sputnik France (1), sharp7272 (1), Institut des Libertés - IDL (1), A D (1), Nain deux trois (1), Parti Anti Sioniste (1), Zehnor (1), REPORTAGE REPLAY (1), Daniel Dupont (1), Tout Simplement Ilian (1), Louis (1), Lettre Afrance (1), Ça Zap Tweet ! (1), Bruno Le Salé (1), jordanix (1), Le Précepteur (1), Joachim Veliocas (1), chebkhaled Lemarocain (1)
Gilets jaunes
Brest Buzz (558), Gilets Jaunes 24/24 7/7 Relai Info (546), Éveil Global Conscience Gilets Jaunes (292), Micka R-P 2.0 (215), Verdi (141), GiletsJaunes Constituants (141), Buzz 2.0 (119), Reporter Youtube (104), N E W s H ALL iens 365 (86), hommes libres (78), Reservoir Apps (71), Line Press (70), Le Globe 777 PARISHAV-HL LarbiHavera (67), Politique France (64), bjp abadie (63), Alix a Toujours Raison VS Alixator Officiel (55), Canal Fi (54), La Famille Jaune gilets jaunes (49), Audiovisuel Créativ'Arts (49), Christophe Ambrosino (48), Family N'Orny (47), Gilet Jaune Blog (41), Alliance Jaune, la révolte par le vote (40), Cemil Choses A Te Dire (40), Gilets Jaunes Commercy (38), la france en colere eric drouet (38), Le Parti des Pas de Parti - Officiel (34), Djemadine (34), Info libre (32), Demo Sophie (31), Média 25 (31), La chaine Terrienne (30), aminegociateur (28), Ramous (27), GVS TV (27), SANGLIER JAUNE (26), Isadora Duncan (26), Demos Kratos (26), MOUVEMENT ESPERANCE (25), Lopez Frédéric (21), Eric Plisson (20), Le Croissant de lumière (19), NOUS VOULONS VIVRE ! (18), Moteur En Colère (18), Christo s (16), on s'ennuie (14), ernesto deupoinzero (14), Dada Bens Gilets jaunes/Infos/musique/débats (13), RÉFRACTAIRE TV (13), sam zirab (12), TV JE SUIS BLOCUS (11), Gilets Jaunes Portail Collaboratif (11), Conseil pour le Maintien Des Occupations (11), Julien LAMOURETTE (11), La MINUTE Gilet Jaune (10), Jean-Claude Meyer (9), Street Politics (9), Mehdi tous est rien (9), Le Média Pour Tous – Officiel (8), site de Jacques Chevalier (8), Infos Dijon (7), Momo Rillon (7), independenzawebtv (7), MrMumudu56 (6), Daniel CELLERIER (6), Hors-Série (5), Scalpa Libre (5), bilbu bilou (5), Ben & Sav - Underground News (5), Lacoste28 (5), un tout petit Grain de Sable (5), ici le peuple (5), VARTELE.fr (5), Chrysalide Post 4 (4), Batu Xaan (4), Lundi Matin (4), Deborah Donnier (4), frenchpatriot1006 (4), L'anti bien pensant (4), NFCA MEDIA (4), Rouen Dans La Rue (4), Valek (4), Born2Heal (4), HORS-ZONE Press (3), chris la parole (3), Yves Barraud (3), ChrysalidePost3 (3), Emotion Pictures (3), louisette lavraie (3), ForcePingouin (3), Passion Animale et végétale (3), MontceauNews (3), Nouvelles Terre (3), Grozeille (3), sambo 82 (3), pierre (2), La Fièvre Jaune (2), Damoclès France (2), j21unis (2), Anonymous 000089 (2), stephane gilet jaune (2), LE RELAYEUR (2), Crayon jaune (2), hasara470 moimeme (2), X X (2), Chouky 7388 (2), Christophe Chalençon Mouvement Citoyen 84 (2), Crysalide Post (2), Casus Vérité (2), Mediacord Productions Vidéo (2), street 45 phénix (2), L'écho du peuple (2), V garou (2), Media Investigation (1), Chrysalide Post (1), Franck FALISE (1), Younes Bouremel (1), Where the Claim is (1), CLIC RIC (1), Charles Demassieux (1), Le Compost (1), DIDIXLOD (1), hubert mlt (1), Georges J. (1), Free Panther (1), benoit dikken (1), vanessa beeley (1), banlieuesrespect (1)
Politique
ATSADA Arezki (250), La Luciole Mélenchantée (154), UPR TV (100), La France insoumise (93), JEAN-LUC MÉLENCHON (91), les Républicains (82), Révolution Permanente (81), Ugo Bernalicis (70), Députée Obono (63), Adrien Quatennens (58),

Journal l'Humanité (45), La France insoumise - Groupe parlementaire (42), Mathilde Panot (40), François Ruffin (37), Groupe MoDem et apparentés (36), Sarah El Haïry (36), Territoires! (36), Benoit Hamon (35), John Doggett (34), PCF - Parti communiste français (34), Laurent Wauquiez (34), Alexis Corbière (29), Lutte ouvrière (29), Jean-Christophe Lagarde (29), LE BON SENS (28), Frédéric Descrozaille (28), Jean-Marie Le Pen (23), Danielle Simonnet (21), UDI (20), Anne Genetet (19), Michelle Tirone (19), Michel Larive (18), François Cocq (18), Clémentine Autain (15), Bastien Lachaud, député insoumis (14), Génération·s (12), Le Député du Jour (12), Muriel Ressiguier (12), Bruno Gollnisch (12), Philippe PASCOT (11), Stéphane Peu (11), Élu-e-s Génération-s Conseil de Paris (11), Leila Chaibi (10), Paul Jorion (10), Députés Socialistes et apparentés (10), Luc Carvounas (10), Eric Coquerel (9), Eric ALAUZET (9), Le Poste (9), Marine Le Pen (9), Les Éveilleurs (9), Régis JUANICO (9), FLORIAN PHILIPPOT (9), HENRY DE LESQUEN (8), Loïc Prud'homme député insoumis (8), Solidarité et Progrès (8), Richard Lioger (8), Jean-Paul Lecoq (7), Député LABARONNE (7), Benoit SIMIAN (7), Dimitri Houbron (6), Ajenwy 92 (6), Sophie Taillé-Polian (6), UDI Agir et Indépendants (6), La République En Marche ! (6), Elsa FAUCILLON, Députée (6), Web FO (6), Fréquence Occident (6), Collège Européen (6), Sylvain Maillard (5), Pierre Dharréville Député (5), La Manif Pour Tous (5), Jean-Luc Lagleize - Député de la Haute-Garonne (5), François Jolivet (5), Licra (5), Aurore Bergé (5), Place publique (5), Jean-Louis MASSON (5), Olivia Gregoire (5), Julien Aubert (4), Élise Fajgeles (4), Patrick Le Hyaric (4), frexit upr (4), Frédéric Petit (4), Marie-Christine Vergiat (4), Catherine Fabre (4), Marie-Ange MAGNE (4), Dino cinieri (4), Didier BAICHERE (4), Permanence de Didier Le Gac (3), europe2019 (3), Thierry Benoit (3), Thomas FERRIER (3), Axel Macky (3), Bilger Philippe (3), Christophe Naegelen (3), Ludovic Mendes (3), collardofficiel (3), Jean-Michel Jacques (3), Sacha Houlié (3), Sylvain Waserman (3), Fiona Lazaar (3), Xavier Roseren (3), Manon Aubry (3), Alexandre Holroyd (2), Laetitia Saint-Paul (2), Paula Forteza (2), Union des Démocrates Musulmans Français (2), Fabien GOUTTEFARDE (2), Sébastien Chenu (2), Marie-Pierre Vieu (2), Laurent Pietraszewski Député du Nord (2), Amelie de Montchalin (2), Xavier Paluszkiewicz, Député de Meurthe-et-Moselle (2), Olivier Dassault (2), Alain Bruneel (2), Olga Givernet (2), Pierre-Yves Bournazel (2), Jean-Paul MATTEI (2), Buon TAN (2), Raphaël SCHELLENBERGER (2), Mohamed Laqhila (2), Marie-France Lorho (2), Joachim Son-Forget (2), Olivier SERVA (2), AmnestyFrance (2), Paul Molac (2), Sophie Panonacle (2), Mouvement pour l'initiative citoyenne (2), Alexandra Louis (2), Xavier Breton (1), Dissidence Française (1), Philippe Michel-Kleisbauer (1), PACE EUROPA (1), Jean Noël Barrot, Député des Yvelines (1), Thomas Gassilloud (1), Jean Francois Mbaye (1), Alliance royale Le parti royaliste (1), Républicains Assemblée nationale (1), Aurélien Pradié (1), Penseur Sauvage (1), Vincent Descoeur (1), Barbara Bessot Ballot, députée de Haute-Saône (1), Typhanie Degois (1), Parti Pirate (1), Philippe Folliot (1), Amis de la Décroissance (1), Aurélien Taché (1), VIGIER Philippe (1), Gabriel Serville (1), Berangere Abba (1), La Ligne Claire (1), les Oubliés de l'Europe (1), Marie-Christine Verdier-Jouclas (1), Les députés communistes (1), Nicolas Turquois (1), Nicolas Dupont-Aignan (1), Elodie Hervé (1), Christian Person (1), Pieyre-Alexandre Anglade (1), Vincent Vauclin (1), André Chassaigne (1), Cédric Villani (1), Adrien Morenas - Député de la 3ème circonscription de Vaucluse (1), Caroline Janvier (1), Eric Ciotti (1), CauseToujours (1)

Information locale

France 3 Nouvelle-Aquitaine (398), France 3 Bourgogne-Franche-Comté (266), France 3 Occitanie (237), France 3 Normandie (167), TV78 - La chaîne des Yvelines (160), France 3 Provence-Alpes Côte d'Azur (151), France 3 Grand Est (146), France 3 Auvergne-Rhône-Alpes (121), France 3 Bretagne (116), France 3 Hauts-de-France (109), France 3 Corse ViaStella (104), Télé Lyon Métropole (100), Sud Ouest (80), France 3 Paris Ile-de-France (66), France 3 Centre-Val de Loire (50), Le Parisien (48), 1ère outre-mer (43), France 3 Pays de la Loire (39), La Provence (38), Mayotte la 1ère (29), Grand Lille TV (24), Martinique la 1ère (21), La Nouvelle République - NRCO (20), Banyulsinfo (18), France 3 Toutes Régions (17), France Bleu (15), Guyane la1ère (15), Paris Normandie (13), tvcarcassonne (12), Kernews (11), Journal La Montagne (9), Tv Languedoc (6), TV Landes (6), RP MEDIAS TV (6), Mayotte officiel (5), dequoioncause (4), Les Nouvelles de Sablé (4), Charente Libre (4), demaintv (4), lemainelibre (3), UNISSONS NOUS (3), Corse-Matin Presse (3), Ouest-France (3), Grand Lyon TV (2), Free Dom (2), Wéo, la télé Hauts-de-France (2), La Commère 43 (2), Journal Citoyen Haute-Marne (2), La Renaissance du Bessin (2), Wallis-et-Futuna la 1ère (1), VAR AZUR (1), DMZ TV (1), Polynésie la 1ère (1)

Médiation

Accropolis Replays (170), Xerfi Canal (65), iReMMO (51), HugoDécrypte (30), Publications Agora (25), lesmardis (12), Jean-Marc Jancovici (10), Changer le monde en 2 heures (9), #Indecis (9), Whip. (8), Loïc Chaigneau - IHT (7), Victor Ferry (7), Weekly Reporter (6), Monkey - l'actu décryptée (6), Esprit Critique (5), La Chronique Politique (5), Sicavonline (5), Nicolas Meyrieux (5), Déclic (5), Mouton Lucide (4), FDMTV - France Diversité Média (4), MONSIEUR FRANTZ (4), lejournaldepersonne (3), Les Choses au Claire (3), Politikon (2), Le Réveilleur (2), Jérémy Brion (2), Les Pensées de Riles (2), Armire (2), Le Dessous des Cartes - ARTE (2), Mr. Sam - Point d'interrogation (2), Majid Oukacha (2), Draw my economy (2), MesFinances (2), Guillaume Deloison (2), 911 AVOCAT (1), le blob, l'extra-média (1), Aude WTFake (1), Autrement (1), 1 jour, 1 question (1), Lettres It Be (1), EXPLICITE (1), Sapiens sur un caillou (1), Atheos (1), Draw my news (1), La Tronche en Biais (1), Accropolis (1), Histony (1), Agriculture et Environnement (1), Stupid Economics (1)

Mainstream

CNEWS (2266), RTL - On a tellement de choses à se dire (2228), Europe 1 (1301), BFMTV (635), RMC (559), Le Figaro (341), 28 minutes - ARTE (338), Public Sénat (236), Télé Matin (190), L'Opinion (182), Boursorama (135), France Culture (121), LeHuffPost (74), 20MinutesFR (69), CLPRESS (67), RADIO RCJ (60), Touche pas à mon poste ! (53), Les Terriens (49), Video investigation (38), franceinfo (38), TVLaTribune (37), Mouv' (33), Les Echos (27), France Inter (26), i24NEWS Français (25), Brut (23), Envoyé Spécial (21), Le Monde (21), Complément d'enquête (20), 6Medias (20), ARTE Radio (17), TV5MONDE Info (14), La Croix (13), ARTE (13), Konbini (12), L'Obs (10), Clique TV (8), Vox Pop - ARTE (7), Le Point (7), Ina Actu (6), Investigations et Enquêtes (6), Géopolitis (6), Radio Nova (6), TV5MONDE (6), MonFinancier (6), L'instant détox (5), Paris Match (3), Telerama (3), Ina Politique (3), L'Express (3), L'Emission politique (2), Actu-Environnement (2), LCI (2), Politis Fr (2), Ina Société (1), Ina Talk Shows (1), Ina.fr Officiel (1), Documentaire Société (1), La Grande Librairie (1), Ben on the road (1)

Médias alternatifs et analyses

Sud Radio (315), Le Média (299), Frédérick Moulin (178), Mediapart (135), Regards (98), Fondation Jean-Jaurès (62), Thinkerview (40), Le Vent Se Lève (29), Christian Mrasilevici (25), Potentia Multitudinis (17), Là-bas si j'y suis (13), Cercle des économistes (10), Sociologie de l'intégration 2 (8), Les Déconnomistes (8), Partager C'est Sympa (7), Le Monde Moderne (6), TARANIS NEWS (6), Guerre de Classe (6), HUB Institute (6), Trouble Fait (6), Fakirpresse (5), Juan Branco (5), Conférence des présidents d'université (5), Break News TV (5), Osons Causer (5), Université Paris-Dauphine (5), Le CERA (4), Acrimed Vidéos (4), le Stagirite (4), Officiel DEFAKATOR (3), Agora Des Savoirs (3), Laïcité République (2), Université Grenoble Alpes (2), disclose ngo (1), Franck Lepage (1), Faculté des Lettres de Sorbonne Université (1), Université Paris13 (1), UnivNantes (1), Jeunes IHEDN (1), UnivParis1 (1), StreetPress (1), 4emesinge (1), François CHILOWICZ (1), Adel Damian (1)