Analyse de groupes Facebook, comparatif lexicométrique des données de crowdtangle à celles accessibles par navigation

Lucie Loubère

Analyse de groupes Facebook, comparatif lexicométrique des données de crowdtangle à celles accessibles par navigation

Lucie Loubere [1]

[1] Université de Toulouse – lucie.loubere@univ-tlse2.fr

Abstract

Social media platforms have become a daily routine for numerous individuals, and the discourses they harbor are expanding year by year. Despite the wealth of data they contain, which serves as a valuable resource for researchers in the humanities, it remains under the ownership and governance of the platforms themselves. In this communication, we scrutinize the CrowdTangle tool, which grants researchers access to data from public Facebook pages and groups, albeit limiting it solely to post content (excluding comments and replies). Our study indicates that while the exportation feature offered by the platform yields more posts than retrieval through browsing, the latter method, by accessing comment content and its responses, unveils specific themes such as information sharing and argumentation.

Keywords: Digital Social Networks ; Facebook ; crowdtangle ; Lexicometry ; Scraping ;Digital Humanities

Résumé

Les réseaux sociaux sont devenus le quotidien de nombreuses personnes, et les discours qu'ils contiennent prennent de l'ampleur d'année en année. Bien que les données qu'ils renferment constituent une ressource pour les chercheurs en sciences humaines, elles restent la propriété des plateformes qui en régissent l'accès. Dans cette communication, nous étudions l'outil CrowdTangle qui permet aux chercheurs d'accéder aux données de pages et groupes Facebook publics, mais en les restreignant aux seuls contenus des posts (sans commentaires ni réponses). Bien que l'exportation proposée par la plateforme renvoie plus de posts que la récupération par navigation, notre étude montre que cette seconde solution, par l'accès aux contenus des commentaires et de leurs réponses, révèle des thématiques spécifiques telles que le partage d'information ou l'argumentation.

Mots clés : Réseaux sociaux numériques ; Facebook ; crowdtangle ; lexicométrie ; scrapping ; humanités numériques

1. Introduction

Les réseaux sociaux numériques (RSN) sont devenus des outils totalement intégrés dans le quotidien des populations. La multiplicité des plateformes et l'accroissement permanent de leurs contenus forment aujourd'hui un réseau de réseau (Granjon, 2017) complexe à appréhender. Qu'ils soient utilisés pour des loisirs, pour des pratiques informationnelles ou encore à destination de mobilisation citoyenne et politique, ils constituent aujourd'hui une agora sur laquelle circulent de nombreux discours. Cette omniprésence des outils de dialogues numériques, leur facilité d'accès à tout un chacun et leurs algorithmes d'affichage de contenus en font des lieux d'échanges cristallisant les polarisations par les effets de bulles de filtres et chambre d'échos (Pariser, 2011). Les discours présents dans ces outils peuvent s'inscrire dans les recherches en sciences humaines, mais sont soumis à des modèles économiques et des licences propriétaires qui régissent l'accès à ces données et *in fine* peuvent orienter les recherches sur ces terrains.

En effet, la multiplicité d'usage de ces outils est à l'image de la multiplicité des dispositifs, mais également des populations les utilisant. Ainsi, plusieurs études ont montré la pertinence du réseau social Twitter (aujourd'hui X) pour l'étude des polémiques politiques (Ratinaud, Smyranios 2016). L'accès aux données que proposait la plateforme par son API facilitée par des outils ne nécessitant pas de programmation (TMICAT, 4CAT) invitait les chercheurs à étudier ce terrain rendu accessible. Cependant, les études se restreignant à cette plateforme ne portent que sur une catégorie de personne spécifique, délaissant d'autres catégories socioprofessionnelles sous représentées, mais utilisant d'autres outils.

Au-delà du choix de la plateforme étudiée, les outils mis à dispositions pour accéder aux données conditionnent également les recherches. Ainsi la plateforme Facebook propose depuis 2020 l'accès pour les chercheurs et journalistes à certaines données. L'interface permet sans avoir de notions de programmation de rechercher des contenus et/ou collecter des posts de pages et groupes publics. Les données accessibles sont limitées en quantité, mais aussi en type d'information. Si certains de ces freins s'expliquent par les législations comme RGPD (anonymisation des posts par exemple), d'autres comme celui de ne pas fournir les commentaires aux posts n'ont pas d'arguments officiels. Cet outil, comme les autres outils d'exports proposés par de tierces applications, dicte donc les données accessibles et par là oriente les travaux de recherches menées.

Notre présentation cherche à poser une réflexion sur l'incidence de l'outil crowdtangle, sur les recherches et de façon plus générale, de restreindre les études des groupes Facebook aux seuls contenus de leurs posts. Pour cela nous proposerons une comparaison quantitative des posts récoltés par crowdtangle, à ceux récoltés par navigation sur les pages des groupes. Nous étudierons également la proportion de posts, commentaires et réponses[1] dans les données recueillies par navigation. La seconde partie de notre travail se focalisera sur l'exploration lexicométrique de corpus composés de posts, commentaires et réponses associées afin d'observer les discours spécifiques à ces types de productions.

2. Méthodologie

2.1 Les corpus

Afin d'explorer les variations d'accès aux données entre les exports de crowdtangle et les données accessibles par navigation, nous avons étudié 3 groupes Facebook publics, qui seront présentés ici par ordre croissant de nombre d'abonnés.

Stop à la 5G France : avec plus de 8000 membres, ce groupe rassemble des opposants au déploiement de la 5G en France. Il existe depuis mars 2019 et son objectif est de « recueillir les informations, ressentis et peurs sur le sujet » des membres.

Le secret de la vérité : ce groupe Facebook comptait plus de 10300 personnes lors de l'export du corpus. Créé en septembre 2021, il revendique de poser les questions sur différents sujets et amener les gens à se réveiller face aux mensonges établis dans la société et/ou maintenus par les élites. Nous y retrouvons donc plusieurs théories du complot (terre plate, mouvance Qanon…).

Zététique : ce groupe est composé de plus de 30 200 comptes et existe depuis 2007. La charte publiée par le groupe présente ce dernier comme un lieu d'échange et de pratique de la

[1] L'interface Facebook permet de commenter les posts, mais aussi de répondre aux commentaires déjà émis sur un post.

Zététique, c'est-à-dire la mise en débat de sujet pouvant être traité de façon scientifique, mais étant affecté par des mythes ou biais suffisamment répandu.

2.1.1 crowdtangle vs scraping

Pour chacun des groupes, nous avons procédé à deux exports :

l'export par scraping : à l'aide de scripts en Python et de la bibliothèque Selenium, nous avons programmé un navigateur afin de faire défiler les posts de chaque groupe par ordre chronologique. Nous avons, à chaque nouveau post rencontré, conservé son identifiant et la date de création du post.

L'export par crowdtangle : la plateforme proposée par Facebook permet d'accéder aux posts de groupes publics et après recherche, d'en exporter les contenus langagiers et autres données au format CSV. Les limitations de données par export nécessitent le découpage des requêtes par temporalité. Ici nous avons effectué des exports mensuels sur chaque groupe.

Afin de pouvoir comparer les deux exports, nous avons sélectionné une plage temporelle présente dans le corpus scrapé et effectué les requêtes sur cette période dans crowdtangle.

	que crowdtangle	que navigation	les deux
5G	183	22	403
Vérité	239	64	298

Tableau 1 : nombre de posts en fonction de l'export.

La comparaison des deux modes d'export nous montre une forte disparité dans les posts récupérés. La méthode par navigation se montre incomplète, allant jusqu'à une différence de 35 % du corpus sur le corpus sur Le secret de la Vérité. Cependant nous notons également la présence de posts dans la méthode par navigation qui ne sont pas proposés par crowdtangle.

2.1.2 Au-delà des posts, les autres contenus des groupes

Pour chaque groupe et après avoir récolté les numéros d'identification des posts, nous avons, pour chaque post, récolté l'intégralité des textes disponibles (hors texte mis en image) des posts, des commentaires puis les réponses à ces commentaires. Ces données ont été enregistrées dans un corpus textuel identifiant pour chaque contribution l'identifiant du post sur lequel étaient publiés les commentaires et réponses, le type d'écrit (post, commentaire et réponse).

5G : Nous avons récolté les contenus associés à 1091 posts (dont 525 contenant du texte, les autres ne contenant que des images ou vidéos ne sont pas inclus dans le corpus).

Vérité: Nous avons récolté sur ce groupe les contenus associés à 393 posts (dont 262 avec texte).

Zététique : Nous avons récolté les contenus associés à 1235 posts (dont 1212 avec du texte).

Les chiffres de cette récolte sont repris dans le tableau 2, ainsi que les pourcentages cumulés dans l'illustration 1. Nous pouvons observer que le les posts représentent moins de 4 % de chaque corpus, les commentaires sont eux dans une fourchette allant de 20 à 30 % et enfin les réponses à commentaires sont elles majoritaire quelque soit le corpus (entre 66 et 75%).

	Nombre de Posts	Nombre de commentaires	Nombre de réponses
5G	525	3859	10716
Vérité	262	1517	5459
Zetetique	1212	17314	31011

Tableau 2 : fréquence des posts, commentaires et réponses dans chaque groupe.

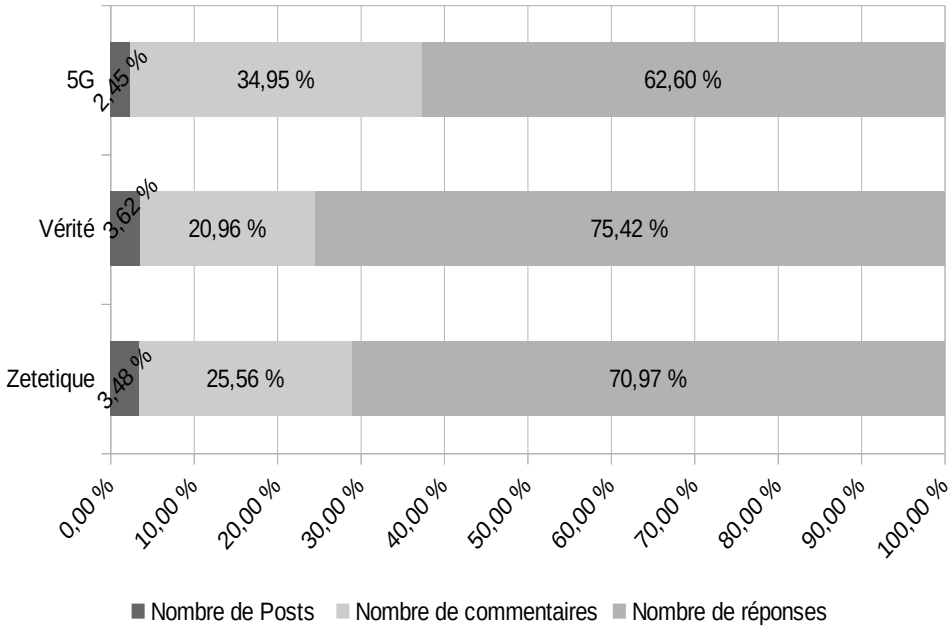

Illustration 1 : pourcentages des posts, commentaires et réponses dans chaque groupe.

2.2 Les outils lexicométriques

Dans notre étude nous mobiliserons plusieurs outils lexicométriques, que nous présentons ici très brièvement.

La classification hiérarchique descendante de type Reinert : initialement développée par Max Reinert (1983), nous utilisons ici la variante proposée par le logiciel Iramuteq (Ratinaud, 2014). Le principe de cet outil est dans un premier temps de segmenter des textes (ici post, commentaires et réponses) en unités de tailles homogènes (environ 35 formes par segment). Une fois ces unités mises à jour, elles permettront de construire un tableau lexical reprenant la présence ou absence des formes actives les plus présentes (verbe, nom, adjectif, adverbe) dans les segments de textes. Ce tableau sera soumis à une analyse factorielle par bipartitions successives afin de regrouper les segments ayant tendance à porter le même lexique. Ces découpages permettent d'accéder aux mondes lexicaux (Reinert, 1993) qui considèrent que les énoncés du discours sont des points de vue à l'interaction du sujet énonciateur, mais aussi de son activité ou de son contexte.

Les analyses de spécificités : ces outils reposent sur un tableau de contingence croisant les formes actives (nom, adjectif, verbe, adverbe) et les modalités d'une variable. À partir de ce tableau sera calculé un indice de spécificité qui prendra la forme d'une valeur positive pour les formes les plus sur-représentées dans une modalité de variable et négative pour une sous-

représentation, la valeur de l'indice rend compte de la corrélation. L'objectif de cette analyse est de mettre à jour le lexique spécifique à un groupe de texte (Lebart et al, 2019).

3. Les analyses

Dans les parties qui suivront, nous n'utiliserons plus les données issues de crowdtangle, mais uniquement celles provenant du scraping. En effet, nous souhaiterions ici mettre en évidence l'apport des commentaires et des réponses dans une étude sur les discours véhiculés dans un groupe Facebook.

3.1. les classifications, des thématiques différentes

Notre démarche est de soumettre nos trois corpus issus de la navigation à des CHD de type Reinert, afin d'observer si les thématiques structurantes des discours présents dans les corpus sont réparties de façon homogène entre les posts et les autres formes d'interactions.

3.1.1 Le groupe Stop 5G France

Ces résultats sont le fruit d'une classification en 65 classes et un seuil de 300 ST par classe, ils reprennent 79 % du corpus en 7 classes.

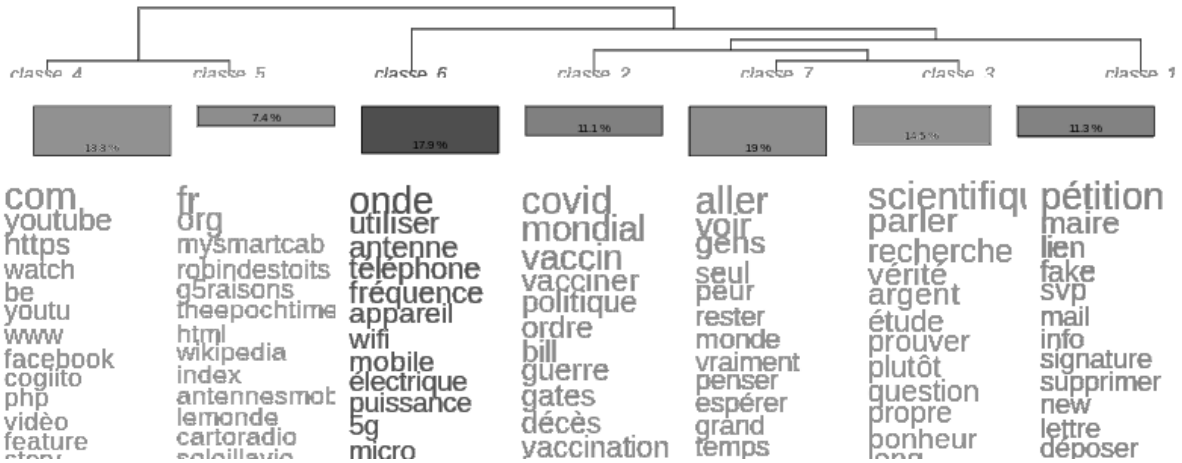

Illustration 4 : corrélation entre le type de segment et la classe sur le corpus Stop 5G France

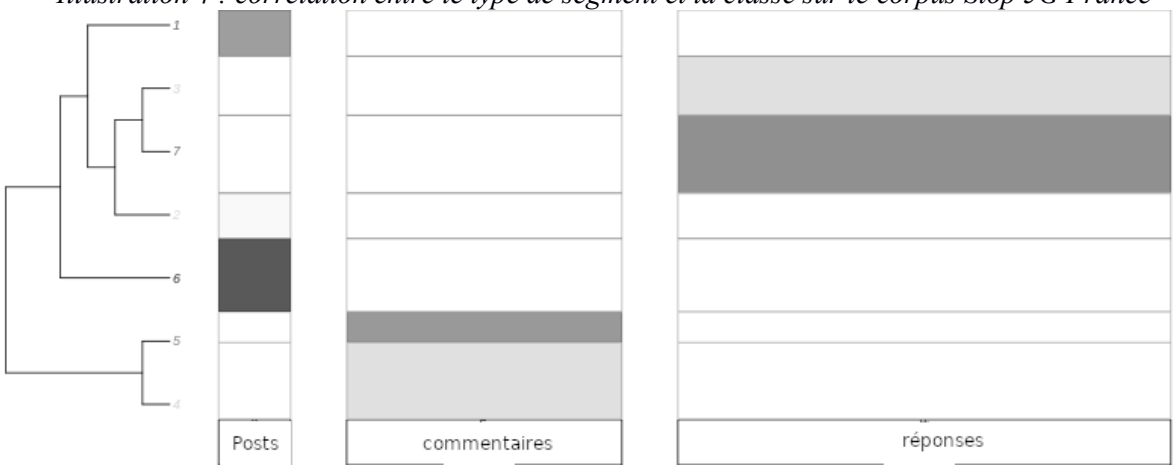

Illustration 5 : corrélation entre le type de segment et la classe sur le corpus Le Secret de la Vérité

Le graphe des corrélations (Ratinaud, 2014) entre type de ST et classe fait apparaître une concentration des ST provenant des posts dans les classes 1 et 6. Elles représentent le discours général des associations de lutte contre la 5G, nous y retrouvons l'information basique sur la 5G et les moyens de lutter contre son développement (pétition, recours des mairies…). Nous notons également dans une moindre mesure (chi2=5) une sur-représentation dans la classe 2 qui porte sur la santé. Dans cette classe, la thématique et les mots portant sur le covid (covid, vaccin, virus…) sont classés très haut dans les profils. Cependant ce discours n'est qu'une faible proportion de la classe (covid n'apparaît que dans 7 % de la classe), où nous observons plus généralement un rapprochement entre les ondes de la 5G et des dégâts sur la santé en général :

> *Segment de la classe 2*
>
> *bon ben voilà c est tout ah oui médite pour l esprit exercice pour le physique mange sainement jeûne pour les défenses mes 4 principes pour m adapter aux ondes qui vont augmenter sous peu onde virus virus vivant ou mort mon intention à moi est simplement la vie ter*

Les commentaires sont sur-représentés dans les classes 5 et 4 qui sont des classes contenant les hyperliens. Enfin, les réponses sont sur-représentées dans les classes 7 et 3 portant respectivement sur les façons de s'informer et l'argumentation sur les points de vue non partagés.

3.1.2. le groupe Le Secret de la vérité

L'analyse que nous présentons ici est le résultat d'une classification en 70 classes avec un minimum de 350 ST. Elle reprend 68 % du corpus en 6 classes

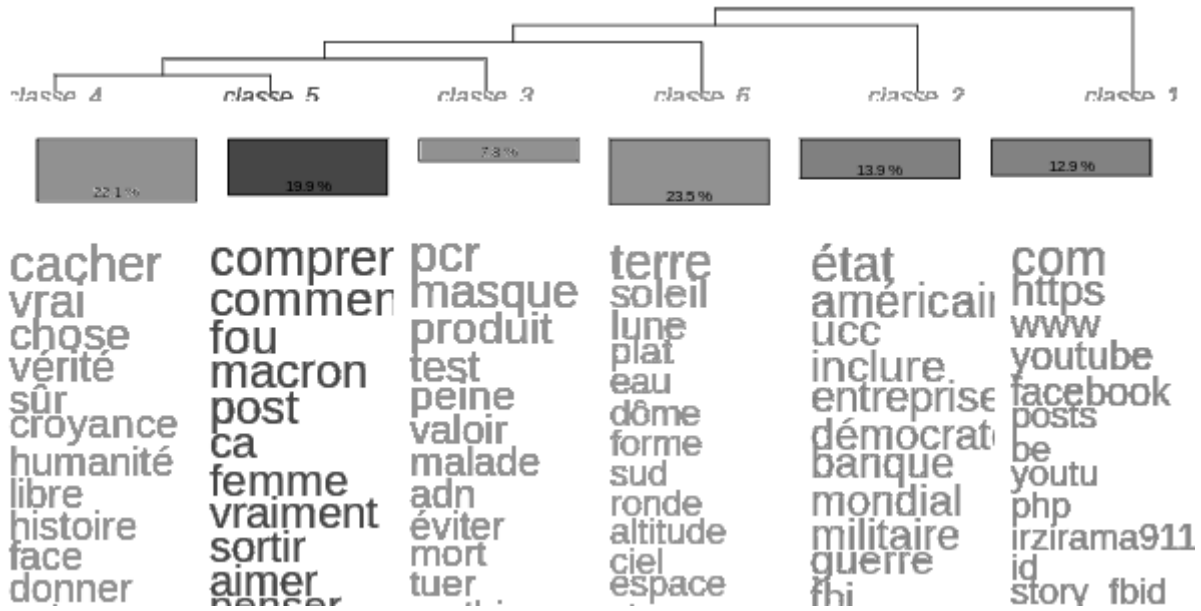

illustration 2 : dendrogramme et profils de la classification sur le corpus du groupe Le Secret de la Vérité

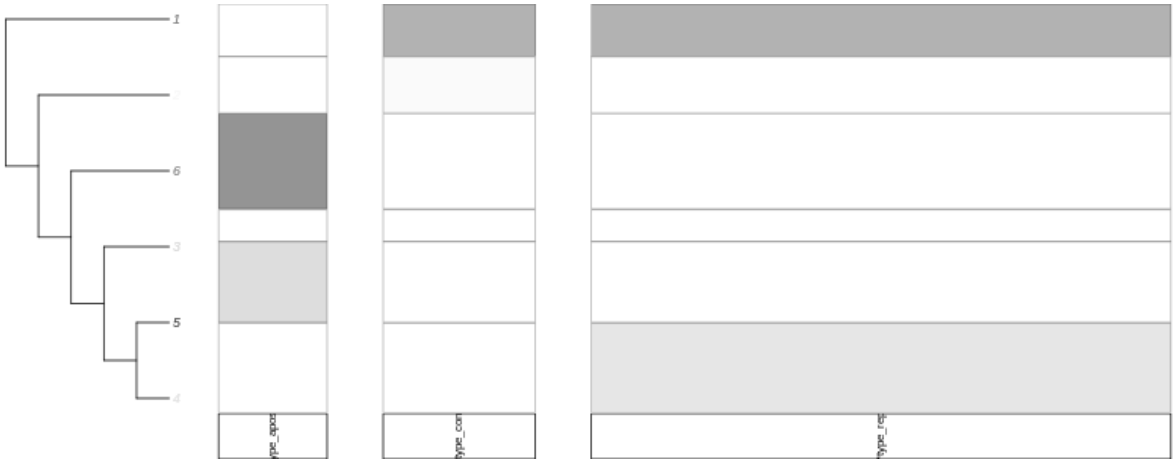

Illustration 3 : corrélation entre le type de segment et la classe sur le corpus Le Secret de la Vérité

La répartition des types de discours nous montre une sur-représentation des segments provenant des posts dans les classes 5 et 6 portant respectivement sur l'actualité politique française et la théorie de la terre plate. Les commentaires sont sur-représentés dans les classes 2 et 1 qui portent sur l'actualité internationale et sur les liens hypertextes (partage d'information supplémentaire). Enfin, les ST provenant des réponses aux commentaires sont surreprésentés sur la classe 1 contenant les liens hypertextes et la classe 4 qui elle porte le discours émancipateur de l'accès à la vérité.

3.1.3 Le groupe Zététique

Cette analyse est le résultat d'une classification en 35 classes avec un seuil de 1400 ST par classe, elle reprend 93 % du corpus en 7 classes.

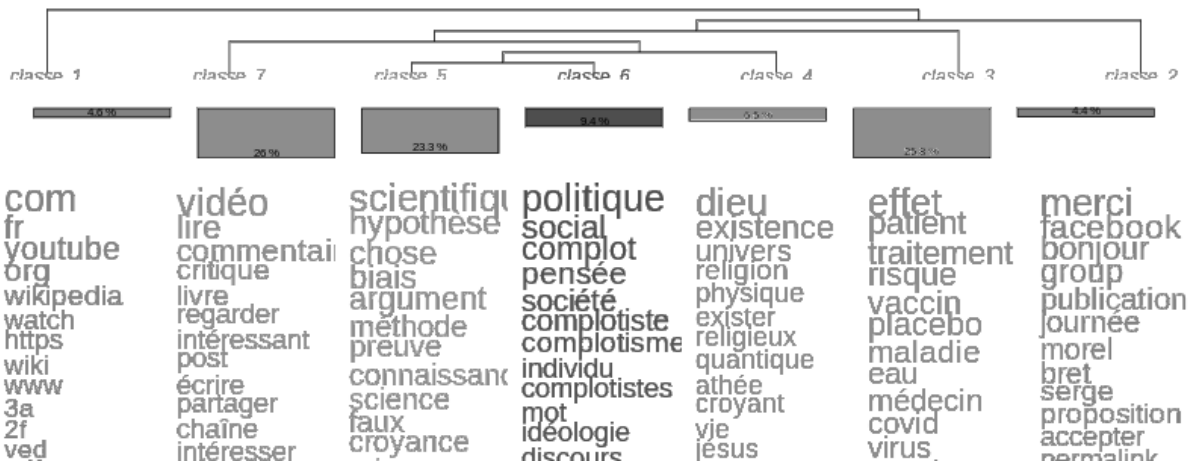

Illustration 6 : dendrogramme et profils de la classification sur le corpus du groupe Zététique

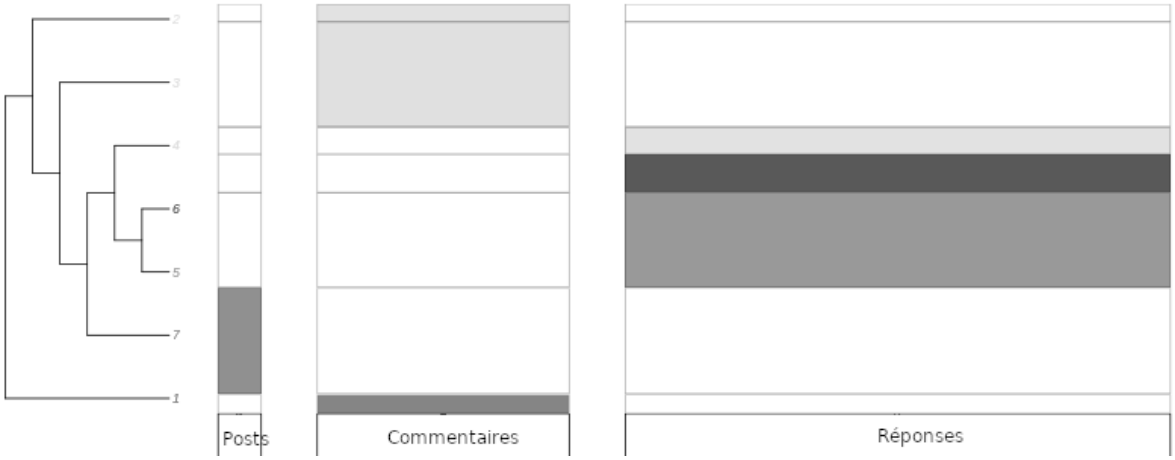

Illustration 7 : corrélation entre le type de segment et la classe sur le corpus Zététique

Nous pouvons observer une répartition des ST présents sur les posts dans la classe 7 qui est centrée sur le partage d'information et/ou de sujet dans la thématique du groupe. Dans cette dernière, les ST des posts y sont sur-représentés mais minoritaires. En effet ils ne représentent que 7 % de la classe (mais plus de 1/3 des ST des posts sont concentrés ici). Cette classe porte principalement sur le partage de sources et la modération du site (qui renvoie sur des sujets déjà traités dans d'autres posts). Il est à noter qu'une grande part des ST provenant de post sont des demandes de sources :

Segment de la classe 1 type post

bonjour à tous je vie au usa et je me demandais si quelqu_un connaissait une vidéo ou documentaire présentant la zététique en anglais ou avec des sous titre j ai cherché chez sur youtube et n ai trouvé jusqu_à maintenant aucune vidéo sur le sujet en anglais merci

Les ST provenant des commentaires sont surreprésentés dans la classe portant sur le domaine de la santé et spécifiquement sur les médecines alternatives. Ici il est question de l'effet placebo avec les notions de bénéfices/risques. Les commentaires sont également surreprésentés dans la classe 1 qui contient les liens hypertextes et enfin dans la classe 2 qui est constituée de formules de politesse.

Pour finir, les ST provenant des réponses se retrouvent spécifiquement dans la classe 4 (qui porte sur les religions), la classe 5 (qui elle est orientée sur des discours spécifiques de la preuve et du raisonnement scientifiques) et enfin la classe 6 (qui questionne les modalités d'adhésion aux théories du complot ainsi que les stratégies pour déjouer ces raisonnements).

3.2. Les analyses de spécificités

Nous avons soumis les verbes des trois corpus à des analyses de spécificités. Bien que le corpus sur la vérité n'ait pas apporté d'informations supplémentaires, nous observons dans les deux autres corpus que les réponses portent de façon spécifique sur des verbes relevant de l'argumentation, tels que **comprendre**, **affirmer**, **confondre**, **parler**, **manipuler**... Cette présence dans les deux corpus pourrait être le signe d'interactions cherchant à contredire un discours.

Tableau 3 : score de spécificités sur les verbes pour les corpus du groupe Stop 5G France (gauche)

	Posts	Commentaires	Réponses
parler	**-1,8041**	**-2,7899**	**4,6955**
balancer	**-0,7758**	**-2,0757**	**3,1682**
aller	-7,9469	0,5168	2,7121
manipuler	**-1,5521**	**-1,2317**	**2,6041**
avancer	**-0,4814**	**-2,2858**	**2,6041**
comprendre	**-1,585**	**-0,9406**	**2,0788**

	Posts	Commentaires	Réponses
comprendre	**-14,477**	**-4,9958**	**14,0304**
tarir	-1,4241	-8,0139	10,0303
marier	-1,0178	-8,7843	9,75
parler	**1,0684**	**-11,4835**	**8,4407**
excuser	-2,5254	-5,4018	7,9268
affirmer	**-3,6536**	**-4,55**	**7,7342**
confondre	**-4,057**	**-2,8888**	**5,6169**
empêcher	-3,5227	-2,6598	5,1196
sentir	-9,0421	-1,1504	4,8737
impliquer	-1,5296	-3,4021	4,7753
considérer	**-3,0385**	**-2,5502**	**4,7607**
relire	**-0,9989**	**-3,5157**	**4,4375**
définir	-3,6952	-2,0675	4,3785
voir	-7,8349	-0,9972	4,2694
répondre	-1,1713	-3,2383	4,2225
rester	-3,6191	-1,5567	3,6386
penser	-4,9225	-1,1788	3,5756
avouer	-0,9667	-2,7618	3,5471
nier	-3,1943	-1,5801	3,3682
aller	-6,5325	-0,7312	3,2189
remarquer	**-0,5769**	**-2,6739**	**2,9858**
relever	**0,2884**	**-3,269**	**2,9677**
imposer	**-3,3321**	**-1,251**	**2,9228**
régir	-1,0616	-1,9353	2,8939
suffire	-4,3831	-0,9759	2,885
démontrer	**-4,0864**	**-0,97**	**2,8042**

et Zététique (droite).

4. Conclusion

L'étude exploratoire que nous avons proposée ici nous a permis de mettre en évidence l'utilité de l'outil crowdtangle pour accéder aux posts publics. En effet il permet de récolter plus de posts que la navigation sur les groupes et son interface ne nécessite pas de compétence de programmation. Cependant un nombre de posts, faible certes, mais existant manque dans ces exports alors qu'ils sont présents par navigation et l'impossibilité d'accès aux commentaires ainsi qu'à leurs réponses reste un angle mort.

La seconde partie de notre exposé interrogeait ces discours manquant dans les possibilités de crowdtangle, à savoir les commentaires et les réponses associées. Nous avons pu observer à partir de 3 corpus de groupes Facebook différents que chaque type de contenu fait émerger des thématiques différentes et a un lexique spécifique. Dans l'ensemble des corpus, nous trouvons de façon sur-représentés les contenus des posts dans un discours officiel et/ou de modération, les commentaires sont systématiquement associés à des liens hypertextes (donc source d'informations supplémentaires) et les réponses se trouvent dans les registres d'argumentation.

La conclusion de cet exemple serait une voie mixte qui exporterait les posts proposés par crowdtangle (qui sont plus nombreux) pour aller récolter par la suite les commentaires et réponses qui y sont associés. Ces résultats ne constituent pas une étude info-communicationnelle à part entière, mais nous amènent à prendre du recul sur les études possibles à partir des exports de crowdtangle et d'une façon plus générale des limites d'accès et de représentativités des données retournées par les outils proposés par les plateformes.

Bibliographie

Granjon, F. (2017). Mobilisations numériques : Politiques du conflit et technologies médiatiques. *Presses des Mines* via OpenEdition.

Lebart, L., Pincemin, B., & Poudat, C. (2019). *Analyse des données textuelles*. PUQ

Ratinaud, P., et Smyrnaios, N. (2016). La Web Sphère De# CharlieHebdo : Une Analyse Des Réseaux Et Des Discours Sur Twitter Autour D'Une Controverse Politique. *ESSACHESS-Journal for Communication Studies*, 9(2), 18.

Reinert, Max. (1983). Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte. *Les cahiers de l'analyse des données*, 8(2), 187-198.

Reinert, M. (1993). Les «mondes lexicaux» et leur «logique» à travers l'analyse statistique d'un corpus de récits de cauchemars. *Langage et société*, 66, 5-39

Pariser, E. (2011). The filter bubble : What the Internet is hiding from you. Penguin UK.

Ratinaud, P. (2014). IRAMUTEQ : Interface de R pour les Analyses Multidimensionnelles de TExtes et de Questionnaires. Consulté à l'adresse http://www.iramuteq.org

Ratinaud, P. (2014) Visualisation chronologique des analyses ALCESTE : application à Twitter avec l'exemple du hashtag #mariagepourtous. Dans : *Actes des 12eme Journées internationales d'Analyse statistique des Données Textuelles (JADT 2014)*. Paris, France, 2014, p. 553-565.