Analyse de réseaux sociaux : vers une pensée intra-communautaire unique ?

Mathilde BOLTENHAGEN, Arnaud DIDES

Janvier-Mai 2017

Résumé : Nous montrons à travers ce projet comment l'étude de réseaux sociaux, ici Twitter, peut amener à une distinction de groupes d'individus. Le projet est basé sur le contexte des élections présidentielles françaises qui nous permettent d'avoir un repère de groupes d'individus. Le projet se divise en quatre grandes parties que nous avons complétées au fur et à mesure du développement : la création d'un jeu de données, l'analyse des données, la visualisation de celles-ci puis l'interprétation des résultats tirés.

Mots-clés : API Twitter, analyse, visualisation, Latent Dirichlet Allocation, Exceptional Model Mining

1 Introduction

Dans le cadre actuel, nous sommes immergés par les réseaux sociaux de nos ordinateurs aux smartphones sans oublier les tablettes et objets connectés. En 2012, d'après l'INSEE [1], un tiers des français de plus de 15 ans participe à des réseaux sociaux. Ces derniers sont une source de partages entre les individus et laissent croire qu'ils leur permettent un accès vers une large palette d'idées et d'opinions. Cette supposition semble être contraire à la réalité. Les idées sont-elles suffisamment partagées ? Les réseaux-sociaux limitent-ils les informations selon des groupes d'individus ? Nous amènent-ils vers une pensée intra-communautaire unique ?

La politique française étant déjà une catégorisation des individus, nous nous sommes appuyés sur les campagnes électorales de la présidentielle 2017 depuis Twitter afin de visualiser des oppositions et similitudes selon le côté politique. L'objectif de nos résultats est de montrer s'il existe, ou non, des critères spécifiques à des groupes d'individus et la liaison que l'on peut faire vis-à-vis de leur opinion politique.

Le projet a suivi le processus général de toute analyse de données selon le livre *Big Data et machine learning : Manuel du data scientist* de Jean-Luc Raffaëlli et Médéric Morel[1]. Effectivement, nous avons d'abord établi les grandes lignes de notre projet avant de créer un jeu de données d'études. Puis nous avons commencé par analyser les similitudes des followers entre les comptes politiciens afin de distinguer des groupes d'individus. Par la suite, nous avons analysé ces groupes dans le but de montrer leurs distinctions grâce à différentes études des tweets collectés. Nous avons développé des visualisations afin d'interpréter nos résultats et de répondre à la problématique. La figure 1 présente les différentes étapes du projet.

1. https://www.insee.fr/fr/statistiques/1281312

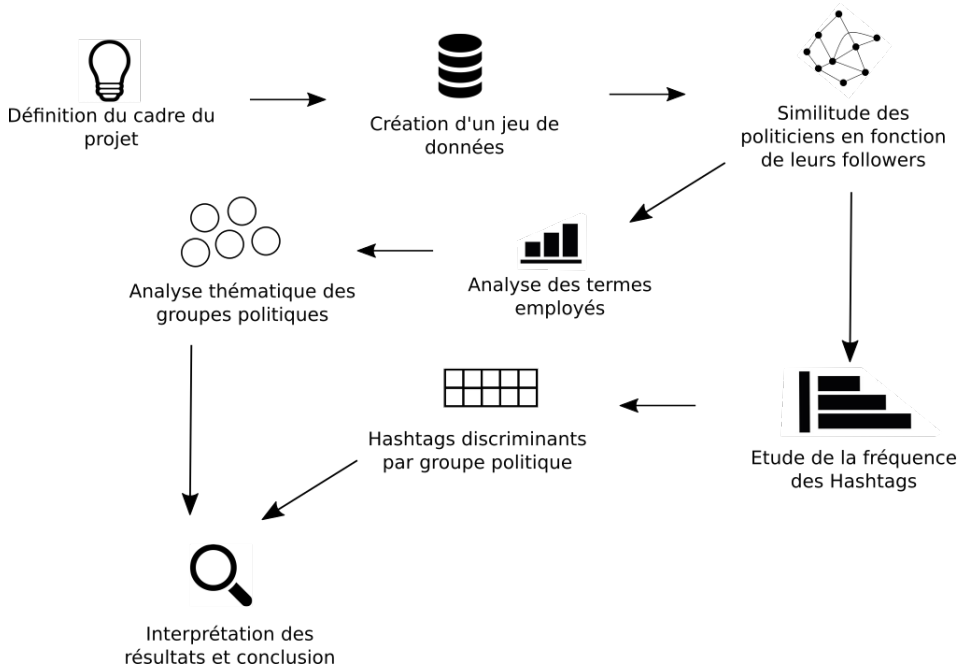

FIGURE 1

2 Création d'un jeu de données d'étude

La première étape de toute analyse de données est la création d'un jeu de données en cohérence avec le besoin de l'étude. Le projet étant basé sur Twitter, nous avons récupéré les informations par le biais de l'API à l'aide de scripts écrits en python a l'aide de la librairie Twython[10].

Le premier jeu de données sur lequel nous nous sommes appuyés est tiré d'une liste de 50 comptes politiques [2] ayant le plus d'abonnés. Nous avons récupérés la liste de leurs followers ainsi que les tweets depuis le 1er novembre 2016. Dans le cadre du projet ANR ContentCheck, les décodeurs proposent une liste de 4500 personnalités politiques françaises avec une colonne indiquant le compte twitter correspondant. Cette liste indiquait un côté politique pour chaque individu ainsi que le parti actuel qui nous servent de repère dans le projet. Ce jeu de données a été mis à jour au cours de notre étude.

3 Analyse et visualisation des données

Suite à la conception du jeu de données, les analyses que nous avons établies ont nécessité une sélection pertinente des informations. Toutes les données n'ont pas utilisées et exploitées de la manière. Nous développerons l'utilisation et les modifications préliminaires des données de chacune des analyses.

L'analyse thématique et l'analyse d'exclusion utilisant respectivement le model Latent Dirichlet allocation[7] et l'Exceptional Model Mining[9] seront plus profondément développées pour leur complexité.

Nos visualisations sont codées en javascript à l'aide de la librairie D3.js[11] qui permet une grande flexibilité sur le rendu.

3.1 Similitudes des followers

Notre première étude s'est portée sur les cinquante premiers politiciens de la liste donnée. Grâce à une comparaison des identifiants de leurs followers, nous avons créé une matrice les connectant entre eux selon

2. http://ymobactus.miaouw.net/labo-top-politiques.php?mode=followers&liste=personnalites

un pourcentage de followers communs basé sur le compte en ayant le moins.

$$similitude_{individu1,individu2} = \frac{\mid followers_{individu1} \cap followers_{individu2} \mid}{min(\mid followers_{individu1} \mid, \mid followers_{individu2} \mid)}$$

La matrice est de la forme (comptes * comptes) avec en guise de poids de l'arc un flottant : le pourcentage de communs décrit précédemment. Ainsi le résultat est normalisé et se représente sous forme de graphe. Le parti politique de chaque compte est noté afin de visualiser si des groupes se créent en fonction du parti.

L'idée de la visualisation est de rapprocher les noeuds ayant beaucoup similitudes. La première idée fût d'utiliser cytoscape.js qui est une librairie open source de visualisation de graphes simple à prendre en main. La distance entre les noeuds est importante dans cette analyse et cytoscape.js ne propose pas une force sur les arcs. Cela permet de déplacer le graphe pour visualiser les noeuds intéressants sans modifier leur valeur. C'est pourquoi nous nous sommes tournés vers D3.js qui propose cette option.

Dans la visualisation (annexe 1) le slider sous le graphe permet de choisir le seuil en pourcentage de followers communs minimal pour une liaison entre deux noeuds représentants les comptes politiques. Par exemple, on peut choisir que les personnalités sont reliées si elles ont au minimum de 60% de followers communs sur la base de celui qui en a le moins.

3.2 Analyse Tropes

Inspirés par l'analyse de Cécile Alduy en 2015 avec le livre *Marine Le Pen prise aux mots, décryptage du nouveau discours frontiste* [8], nous trouvons le logiciel libre Tropes qui permet une analyse sémantique de textes. Le logiciel a été créé en 1994 et se base sur une analyse morphosyntaxique, un lexique et un réseau sémantique qui attribuent une catégorie aux mots selon leur appartenance à un domaine grâce à une l'intelligence artificielle.

Nous avons lancé cette étude à deux reprises lors de notre développement : d'abord pour une analyse des tweets par parti politique parmi les cinquante politiciens les plus populaires donnés, puis sur les onze candidats officiels (annexe 2).

Une visualisation en bâton permet une comparaison lisible entre les résultats. On peut ainsi remarquer rapidement si un des groupes se différencie.

3.3 Utilisation des hashtags

Afin de reconnaître rapidement si certains thèmes sont plus discutés par les personnalités possédant certaines opinions, nous avons créé une visualisation de l'utilisation des hashtags. A partir du jeu de donnés contenant tous les tweets depuis novembre 2016 pour 385 personnes, nous avons récupéré les hashtags afin de déterminer pour chaque jour les hashtags les plus utilisés dans l'ensemble, et leur utilisation pour chaque côté politique.

A partir de ces données nous pouvons afficher un diagramme en barres où chaque barre représente un hashtag et est découpée par couleurs, selon la proportion d'utilisation de ce hashtag par les différentes opinions (annexe 3). Cela permet d'identifier rapidement si certains hashtags sont plus utilisés par certains courants politiques. De plus, il est possible d'ajuster la période représentée ce qui permet de visualiser l'évolution dans le temps. Cependant, cet approche ne permet pas de déterminer automatiquement les hashtags discriminants.

3.4 Thématique des tweets - Latent Dirichlet allocation

L'allocation de Dirichlet latente (LDA) est un model d'allocation probabiliste attribuant des thèmes appelés topics à un ensemble de documents. Très utilisé pour la recommandation d'articles, le model permet de conclure, selon les paramètres donnés, un nombre de termes pour un nombre de thèmes définissant le corps du texte. Cela permet d'avoir une vision des sujets évoqués et ainsi comprendre les centres d'intérêts de l'utilisateur dans le cas de la recommandation d'articles. Dans notre étude, nous allons voir si les idées sont stables ou s'opposent selon l'opinion politique des comptes twitters étudiés.

3.4.1 Explication du modèle

Le modèle LDA est un modèle bayésien à 3 couches : la collection de documents, la liste des topics et les mots du corpus. Le model peut se représenter par le graphe de la figure 2 où les rectangles représentent une répétition sur les noeuds qu'ils entourent. Les différentes variables correspondantes sont :

- α : vecteur des topics
- D : nombre de documents
- θ : document en question
- $z_{d,i}$: assignation des topics du mot i du document d
- $w_{d,i}$: i-ème mot du document d
- N : nombre de mot du document
- ϕ : k-ème topic
- K : nombre de topic
- β : vecteur des mots

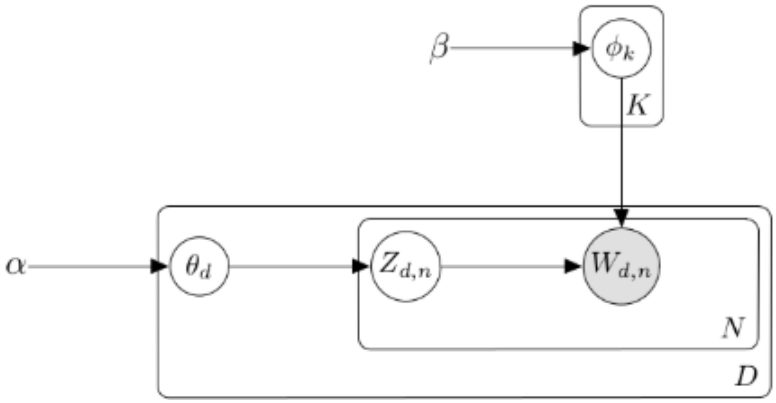

FIGURE 2

La traduction mathématique de ce graphique est définit par :

$$p(w, z, \theta, \phi \mid \alpha, \beta) = p(\theta \mid \alpha)p(\phi \mid \beta)p(z \mid \theta)p(w \mid \phi, z)$$

Le premier facteur du membre de droite représente la distribution des topics par documents. C'est la première étape du modèle LDA. Cette distribution suit la loi Dirichlet qui, avec le paramètre alpha, vaut :

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \theta_1^{\alpha_{k-1}} ... \theta_K^{\alpha_{K-1}}$$

Le paramètre α est un vecteur donnant les poids antérieurs des topics par document. Afin d'avoir une bonne répartition des topics, il est conseillé de mettre un poids inférieur à 1. Le résultat de cette étape donne une proportion des topics par documents. Le tableau ci-dessous en est un exemple illustratif.

	topic 1	topic 2	topic 3
Document 1	0.5	0.2	0.3
Document 2	0.1	0.4	0.5
Document 3	0.1	0.12	0.78

Le second facteur est la distribution des termes des topics. Chaque topic se voit attribué une probabilité par terme et un certain nombre de termes définit par l'utilisateur. Cette distribution suit une fois de plus la loi Dirichlet avec le paramètre β pour la répartition des mots :

$$p(\phi \mid \beta) = \prod_{k=1}^{K} \frac{\Gamma(\beta_k)}{\prod_{v=1}^{V} \Gamma(\beta_{k,v})} \prod_{v=1}^{V} \phi_{k,v}^{\beta_{k,v}-1}$$

	terme 1	terme 2	terme 3	terme 4
topic 1	0.1	0.7	0	0.2
topic 2	0.3	0	0	0.7
topic 3	0.8	0.1	0.1	0

L'étape suivante du modèle LDA donne la probabilité d'un mot d'un document d'appartenir à un topic. Ainsi pour chaque mot de chaque document, un topic est attribué en fonction de la distribution θ du document par la formule suivante.

$$p(z \mid \theta) = \prod_{d=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{n_{d,k}}$$

Voici une bonne visualisation de cette étape :

	mot 1	mot 2	mot 3
Document 1	topic 2	topic 3	topic 2
Document 2	topic 1	topic 3	
Document 3	topic 3	topic 2	topic 4

Le dernier facteur du modèle est un ré-échantillonnage d'un mot w selon la probabilité du topic à générer ce mot :

$$p(w \mid z, \phi) = \prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{n_{k,v}}$$

3.4.2 Mise en pratique du modèle

Le jeu de données de cette analyse contient les tweets de 385 comptes politiques de la liste donnée par Le Monde, leur date et leur orientation politique (Extrême-droite, Droite, Centre, Gauche, Ecolos et Extrême-gauche). L'idée de l'analyse est de sortir les thèmes des six côtés afin de voir leur évolution dans le temps et de les comparer.

Nous avons d'abord essayer d'utiliser l'Allocation Dirichlet Latente en Python avec la librairie Gensim mais le logiciel Knime [13] nous a paru beaucoup plus pertinent pour la maniabilité des données qu'il propose. Par le biais de filtres sur colonnes et lignes, nous avons sorti les thèmes des côtés politiques par période du 1er novembre jusqu'au 30 avril. Nous avons préalablement retiré les mots appelés `stop words` très utilisés et insignifiants dans une analyse thématique tels que les déterminants.

La visualisation de ce travail est en quatre dimensions : le temps, le groupe politique, les thèmes et les mots des thèmes. D3.js propose des visualisations de `pack` comme des bulles dans lesquelles on peut regrouper d'autres bulles (annexe 4). Ainsi chaque bulle représente un groupe politique dans laquelle il y a quatre bulles contenant chacune quatre mots : c'est le résultat du modèle LDA. Sous le graphique, un axe temporel permet de se situer et de faire évoluer les thèmes des groupes dans le temps. Il est aussi possible de zoomer sur un côté politique (annexe 5).

3.5 Subgroup Discovery/Exceptional Model Mining

L'Exceptional Model Mining (EMM) est une technique de data mining développée par Wouter Duivesteijn, Ad J. Feelders et Arno Knobbe. Elle a pour but d'identifier des sous groupes intéressants parmi des données.

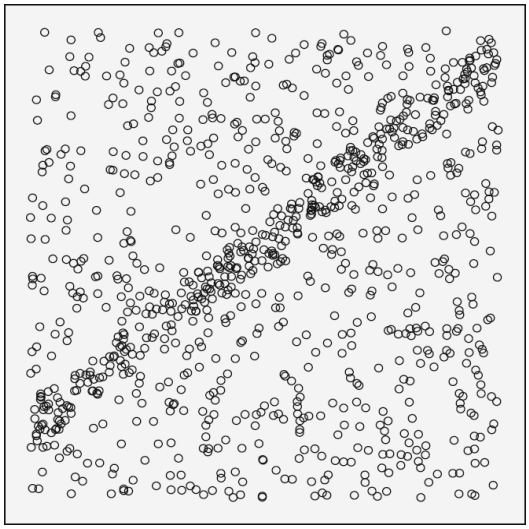

FIGURE 3

Dans cet exemple (figure 3), on possède un jeu de données composé d'attributs cibles (representés sur les deux axes) d'attributs descriptifs non representés. Chaque point possède ainsi un ensemble d'attributs qui le distingue. On observe facilement qu'il existe un sous-groupe de données qui se comporte différemment et forme cette diagonale. Le but de l'EMM est d'identifier une description de ce sous-groupe par rapport aux attributs descriptifs. La figure 4 correspond ainsi aux points qui vérifient cette description tandis que la figure 5 décrit le complément de ce sous-groupe.

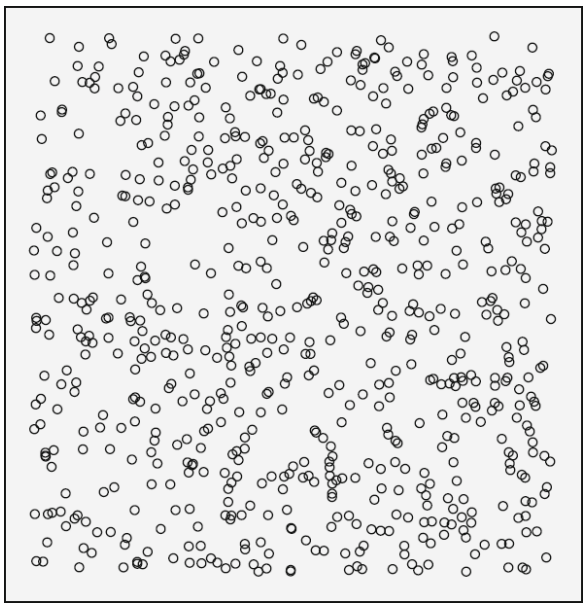

FIGURE 4

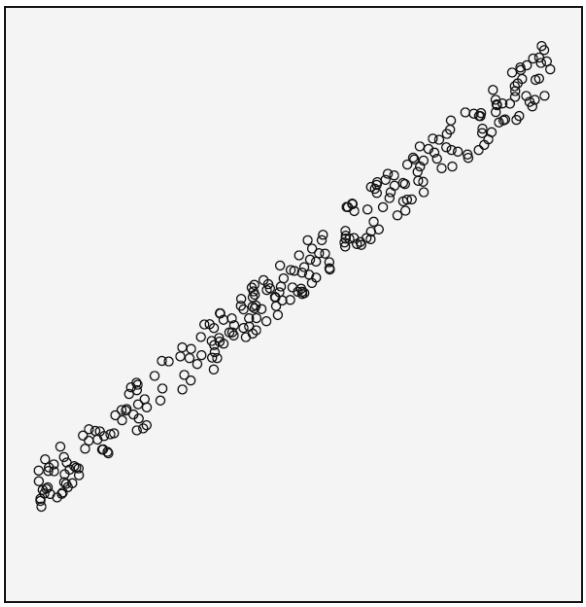

FIGURE 5

Il s'agit d'abord de séparer les données en deux types d'attributs :

— les attributs descriptifs
— les attributs cibles

On cherche alors un sous-groupe des données caractérisé par certaines conditions sur les attributs descriptifs pour lequel les attributs cibles ont un comportement jugé intéressant. On a alors un concept de qualité d'un sous-groupe.

Les critères de qualité pouvant être utilisés sont la différence de distribution d'un attribut cible, la corrélation entre deux attributs cibles mais également la taille du sous-groupe. En effet, il est plus facile d'avoir un sous-groupe avec un comportement différent si il est très petit. Les mesures de qualités utilisés sont souvent une combinaison de ces critères.

Dans notre cas, la mesure de qualité utilisée est la WRAcc (Weighted Relative Accuracy) car elle offre un bon compromis entre taille du sous-groupe et la précision comparée à d'autres mesures telles que la précision qui produit des sous-groupes trop petits, ou l'indice de Jaccard qui trouve des grands sous-groupes mais trop diversifiés.

En plus de la mesure de qualité, il faut choisir un algorithme de recherche des sous-groupes. Nous

avons choisi une recherche en faisceau, qui était conseillée. Cette recherche prend alors en paramètres deux variables, la largeur w et la profondeur d. A chaque niveau, on génère des nouvelles conditions puis calcule la qualité du sous-groupe défini. On ne garde ensuite que les w sous-groupes qui ont la plus grande qualité, puis on les explore en rajoutant à chaque fois une condition. On garde en parallèle une liste limitée des meilleurs sous groupes.

La génération des conditions est différente selon le type d'attributs descriptifs. Pour les booléens, les deux conditions générées sont simples, `booléen = true` et `booléen = false`. Pour les attributs à valeurs nominales, on génère une condition pour chaque valeur possible. Pour les attributs discrets, l'algorithme crée des intervalles qui découpent l'ensemble actuel en ensemble de même taille. Le nombre d'intervalles est prédéfini et paramétrable.

Subgroup discovery peut être considéré comme un cas particulier de l'EMM où on ne vise qu'un seul attribut cible. C'est cette variante que l'on a principalement employée, en utilisant comme attribut cible l'orientation politique (Extrême-droite, Droite, Centre, Gauche, Ecolos, Extrême-gauche). Nous avons utilisé plusieurs ensembles d'attributs descriptifs différents pour cette analyse. Nous sommes partis de l'ensemble des tweets des 385 comptes depuis le premier novembre, que nous avons séparé par intervalle de dates. A partir de cela, nous nous sommes d'abord intéressés aux hashtags utilisés par les différents comptes. Les attributs descriptifs étaient dans ce cas la fréquence relative de chaque hashtag pour chaque personne. Nous nous sommes limités à une profondeur de 1 (un seul hashtag) car dans les cas que nous avons essayé, augmenter la profondeur n'apporte que très peu d'informations intéressantes. Nous avions en effet de nombreux sous-groupes avec un hashtag qui caractérisait fortement le groupe et d'autres qui l'affinaient très légèrement. Nous avons ainsi décidé de ne pas les inclure afin d'éviter les confusions. C'est également pour cette même raison que nous avons limiter les conditions à l'opérateur <=. Les données en entrée se présentent ainsi sous cette forme :

	Côté	hashtag 1	hashtag 2	hashtag 3
Personne 1	Droite	0.1	0.5	0.4
Personne 2	Gauche	0.4	0.2	0.1
Personne 3	Droite	0.0	0.3	0.7
Personne 4	Extrême-Droite	0.2	0.5	0.3
Personne 5	Gauche	0.6	0.2	0.2
Personne 6	Centre	0.4	0.3	0.3
Personne 7	Droite	0.0	0.5	0.5

	Côté	hashtag 1	hashtag 2	hashtag 3
Personne 1	Droite	0.1	0.5	0.4
Personne 2	Gauche	0.4	0.2	0.1
Personne 3	Droite	0.0	0.3	0.7
Personne 4	Extrême-Droite	0.2	0.5	0.3
Personne 5	Gauche	0.6	0.2	0.2
Personne 6	Centre	0.4	0.3	0.3
Personne 7	Droite	0.0	0.5	0.5

Les résultats obtenus se présentent ainsi sous la forme de conditions qui caractérisent un groupe. Dans notre cas, celles-ci sont de la forme `hashtag1 > 0.05`. Si la qualité du sous-groupe est bonne, on peut alors dire que les personnes utilisant beaucoup le hashtag 1 forment un groupe constitué principalement de membres partageant la même opinion politique ciblée. Si par exemple dans le tableau plus haut on a comme cible le côté avec comme valeur `Droite`, on aura en résultat un sous-groupe caractérisé par `hashtag3 >= 0.4`.

	Côté	hashtag 1	hashtag 2	hashtag 3
Personne 1	Droite	0.1	0.5	0.4
Personne 2	Gauche	0.4	0.2	0.1
Personne 3	Droite	0.0	0.3	0.7
Personne 4	Extrême-Droite	0.2	0.5	0.3
Personne 5	Gauche	0.6	0.2	0.2
Personne 6	Centre	0.4	0.3	0.3
Personne 7	Droite	0.0	0.5	0.5

	Côté	hashtag 1	hashtag 2	hashtag 3
Personne 1	Droite	0.1	0.5	0.4
Personne 3	Droite	0.0	0.3	0.7
Personne 7	Droite	0.0	0.5	0.5

Il s'agit ici d'un cas où la qualité du sous groupe est parfaite puisqu'il contient tout les éléments qui possèdent la valeur de l'attribut cible visé.

Pour représenter ces résultats, nous avons choisi un tableau. Les colonnes représentent les intervalles de temps et les lignes sont les opinions politiques. Dans chaque case on a les hashtags présents dans les meilleur résultats. De cette manière, observer une colonne permet de comparer les différentes opinions et observer une ligne permet de voir l'évolution dans le temps. Une fonction de surlignage est présente afin de faciliter ces comparaisons.

L'implémentation de l'EMM utilisée est Cortana [12], un logiciel développée par des chercheurs de l'Université de Leiden. La réalisation de cette analyse s'est basée principalement sur le module Cortana pour Knime [13], qui permet d'automatiser facilement le traitement des données. L'extraction des hashtags à partir des tweets est faite en Python.

3.6 Interprétation des résultats et Conclusion

Les différentes méthodes d'analyse et de visualisation présentées précédemment nous permettent d'établir quelques conclusions en rapport avec le sujet que nous rappelons : *les réseaux sociaux nous amènent-ils vers une pensée intracommunautaire unique ?* Notre travail se divise en deux parties : la première montrant qu'il existe une séparation des individus et la deuxième montrant que des groupes se distinguent.

Dans la première analyse, on peut voir que les comptes politiques d'un même parti ont plus de followers en commun qu'avec les comptes de parti différent. On distingue bien les partis à partir de 60% de followers communs. On comprend par là que les followers auront tendance à suivre les membres d'un parti politique plutôt que des politiciens variés. Plus généralement, cette remarque nous amène à penser que si les followers sont focalisés sur un type de comptes Twitter alors le type d'informations auquel ils sont abonnés sera restreint. On cherche donc à montrer dans la suite de notre développement que les partis politiques se distinguent et ne partagent pas les mêmes données.

L'analyse de l'utilisation des hashtags et l'analyse via le logiciel Tropes sont toutes les deux des analyses statistiques simples montrant des différences d'utilisation de termes dans les tweets des comptes politiques. Par exemple, on remarque que la primaire de la Gauche était très majoritairement discutée par des comptes de Gauche mais que la primaire de la Droite était discutée de manière à peu près égale par la Gauche et la Droite. On remarque également de manière évidente que les partis utilisent principalement les hashtags du type *#Fillon2017* pour parler de leur candidat, tandis que leur opposants se contente de *#Fillon*.

Dans un autre sens, la visualisation du résultat de l'analyse sémantique permet de distinguer les candidats. On remarque, par exemple, que le candidat Philippe Poutou utilise particulièrement des connec-

teurs d'opposition. Inversement il n'utilise pas de connecteur de but : ces deux critères l'opposent à ses adversaires. On note aussi que les 4 principaux candidats n'utilisent pas plus de 7% de leur pronoms la première personne du singulier. Cependant le *nous* occupent 40% des pronoms de ces candidats. Ces différences d'utilisations du langage sont remarquables et les followers sont donc habitués à des informations utilisant certains hashtags ou certaine syntaxe.

L'analyse thématique du modèle LDA nous apporte les sujets des côtés politiques dans le temps. On remarque facilement que l'extrême-droite parle de sécurité et de nationalité avec des mots qui reviennent comme *immigration*, *français* et *islamiste* pour quasiment toutes les périodes. Les termes du groupe Ecolos sont fréquemment le *nucléaire* (5/7périodes) et la *pollution* (4/7périodes). Ces résultats et ceux des autres groupes politiques non-décrits nous affirme totalement la stabilité des sujets des tweets selon l'orientation politique.

De manière similaire à l'analyse thématique, la découverte des hashtags discriminants à l'aide de l'algorithme de Subgroup Discovery permet de mettre en évidence des thèmes discutés par les côtés politiques. On retrouve ainsi un mélange de hashtags qui correspondent au soutien pour un candidat (*#fillon2017*, *#marine2017*, etc), à des attaques sur les autres candidats (*#levraifillon*, *#penelopegate*), à des discussions sur des sujets d'actualité (*#justicepourtheo*, *#trump*) et aussi à des thèmes importants pour les partis (*#migrants*, *#nucléaire*). L'évolution dans le temps nous permet de voir par exemple que le hashtag *#fillon2017* a mis un mois après la primaire avant d'apparaitre pour la Droite tandis que *#hamon2017* est apparu directement après.

Toutes ces analyses confirment notre hypothèse : les individus d'un groupe, ici l'orientation politique, partagent une catégorie spécifique d'informations. Les followers sont distincts selon l'opinion politique et se renferment donc sur cette dite-catégorie partagée. Les réseaux sociaux laissent les utilisateurs choisir leur affinité et ainsi créer des liens à des individus qui leur sont similaires ou s'inscrire à des groupes aux mêmes idéologies. Cette interprétation a été étudié par Hygiène Mentale [3], une chaine Youtube de vulgarisation de l'esprit critique, qui disait dans une de ces conférences que les individus aiment suivre ce qu'ils savent déjà ou ce qu'ils approuvent. Pourtant, il est plus pertinent d'affronter ses idées et se forcer à suivre les actualités que l'on n'adhère pas afin de garder l'esprit ouvert.

3. `https://www.youtube.com/user/fauxsceptique`

Annexes

1 Graphe des followers des 50 comptes politiciens

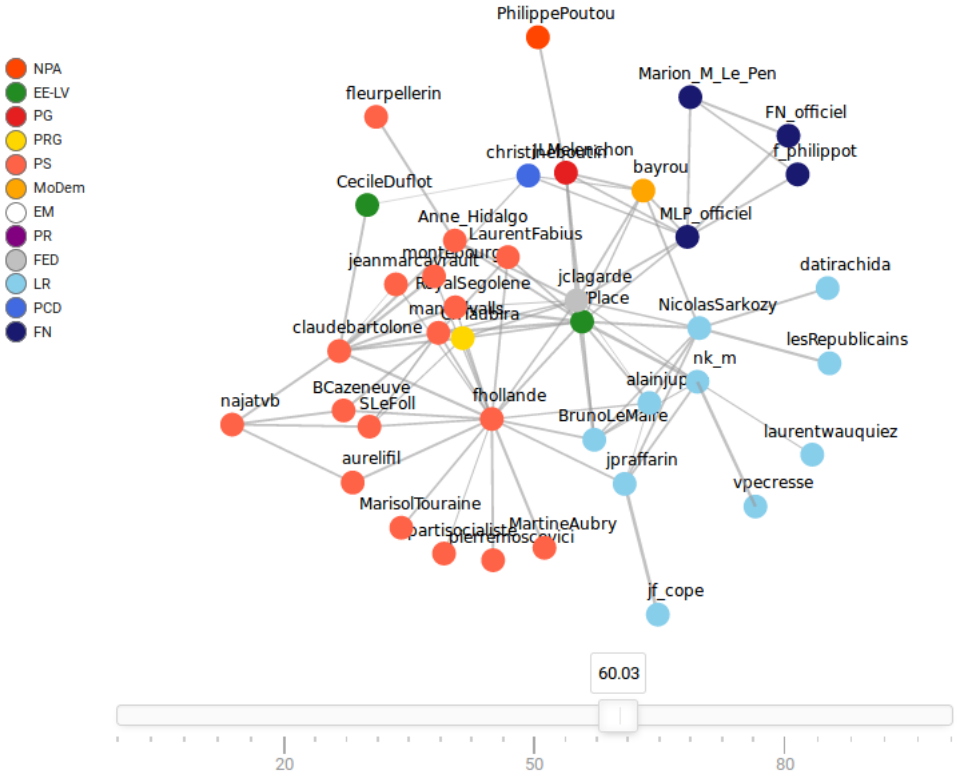

FIGURE 6 – Deux noeuds sont reliés si au moins 60.66% des followers du politicien en ayant le moins sont communs

2 Diagramme syntaxique des tweets avec le logiciel Tropes

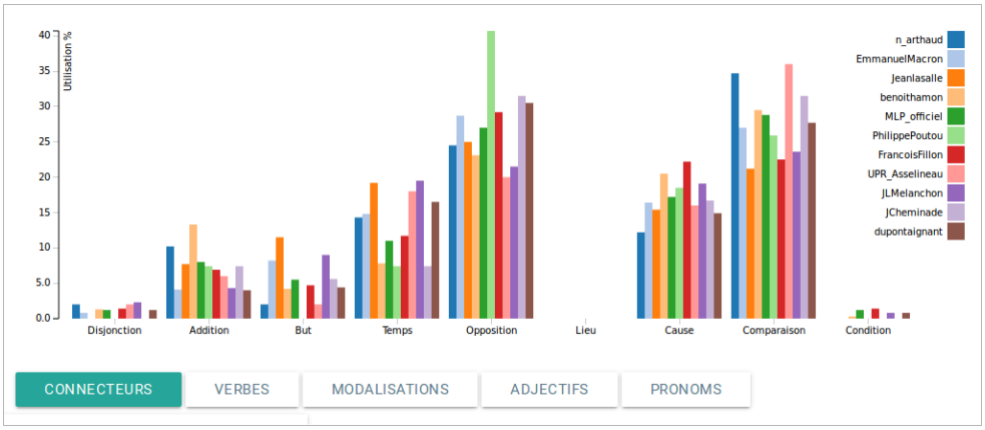

FIGURE 7 – Affichage des connecteurs pour les 11 candidats

3 Répartition des hashtags

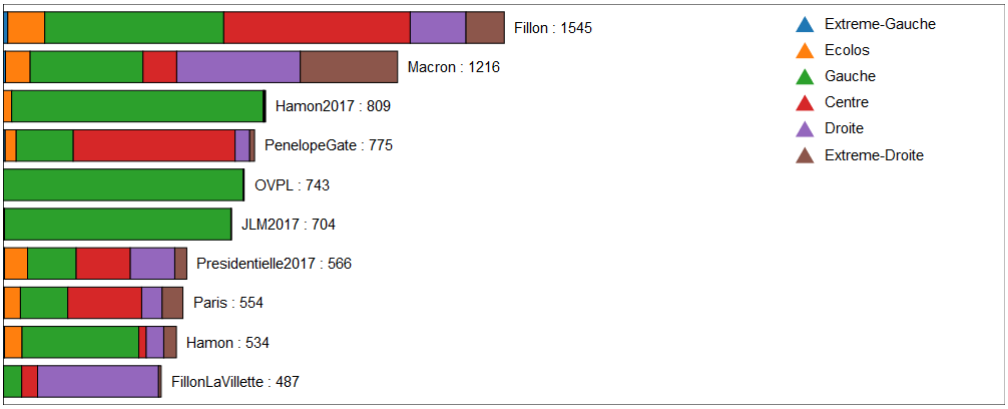

FIGURE 8 – Affichage de 10 hashtags entre février et mars

4 Représentation du résultat de l'algorithme LDA

FIGURE 9 – Affichage de la période début janvier 2017

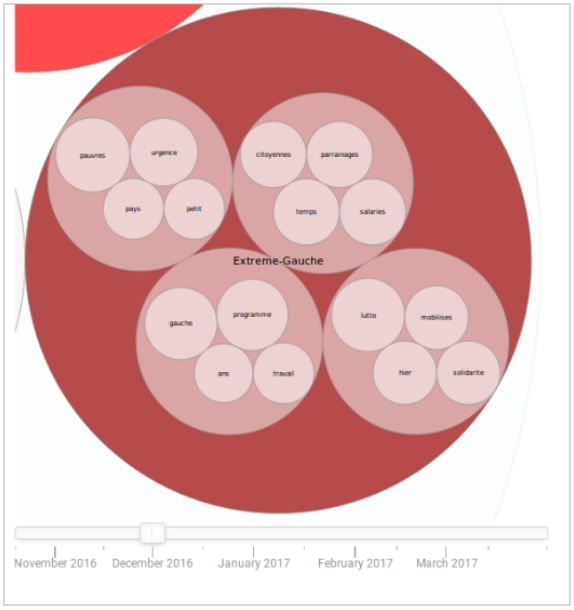

FIGURE 10 – Zoom sur l'extrême-gauche, mois de décembre 2016

5 Résultat du modèle discriminant EMM

	Novembre	Decembre	Janvier	1-14 Fevrier	15-28 Fevrier	1-15 Mars	16-31 Mars
Extreme-Gauche	#trump #presidentielle2017 #fidelcastro	#cahuzac #valls	#fillon	#justicepourtheo #fillon			
Droite	#primaireledebat #primaire2016 #laprimaireledebat	#berlin #hollande #noel	#fillonlavillette #fillon2017 #cnlesrepublicains	#macron #fillon2017 #louvre	#macron #fillon2017 #colonisation	#projetfillon #fillon2017 #trocadero	#fillonpresident #fillon2017 #londres
Gauche	#7novembre16h34 #directan #violencesfaitesauxfemmes	#primairescitoyennes #primaire #directan	#primaireledebat #primairescitoyennes #cpa	#hamon2017 #mutualite #hamon	#portugal #hamon2017 #conseilcitoyen	#hamon2017 #journeedesdroitsdesfemmes #8mars	#hamon2017 #bercy #hamonbercy
Ecolos	#cop22 #nddl #eelv	#climat #pollution #eelv	#faucheursdechaises #fessenheim #enr	#climat #nucleaire #ceta	#jadot #ceta #eelv	#fukushima #8mars #turquie	#trump #climat #repondonspresent
Centre	#udi #laprimaire #trump	#presidentielle2017 #europe #hollande	#ue #trump #presidentielle2017	#presidentielle2017 #penelopegate	#bourdindirect #bayrou #sia2017	#udi #presidentielle2017 #journeedelafemme	#trump #brexit #legislatives2017
Extreme-Droite	#trump #levraifillon #migrants	#noel #fillon #migrants	#trump #marine2017 #migrants	#aunomdupeuple #bobigny #assisesmlp	#liban #marine2017 #mlptf1	#macron #aunomdupeuple #marine2017	#aunomdupeuple #marine2017 #debattf1

FIGURE 11 – Affichage focalisé sur le côté Gauche pour les premières semaines de février

Sources et liens

[1] Big Data et machine learning : Manuel du data scientist, Jean-Luc Raffaëlli, Médéric Morel https://books.google.fr/books/about/Big_Data_et_machine_learning.html?id=IlmbBgAAQBAJ

[2] Introduction to Latent Dirichlet Allocation, Edwin Chen, http://blog.echen.me/2011/08/22/introduction-to-latent-dirichlet-allocation/

[3] Extensions and Adaptations of LDA, Lisa Posch, http://topicmodels.west.uni-koblenz.de/ckling/tmt/part3.pdf

[4] LDA Topic Models, Andrius Knispelis, https://www.youtube.com/watch?v=3mHy4OSyRf0

[5] Introduction aux modèles probabilistes utilisés en Fouille de Données, Théo Francesiaz, Raphaël Graille, Brahim Metahri, http://www-ljk.imag.fr/membres/Marianne.Clausel/Fichiers/Rapport_Metahri_Graille_Francesiaz.pdf

[6] Cours de Natural Language Processing, texte et machine learning, Xavier Dupré, http://www.xavierdupre.fr/app/ensae_teaching_cs/helpsphinx/notebooks/td2a_some_nlp.html

[7] Latent Dirichlet Allocation in R, Martin Ponweiser, http://epub.wu.ac.at/3558/1/main.pdf

[8] Marine Le Pen prise aux mots, décryptage du nouveau discours frontiste, Cécile Alduy, https://dlcl.stanford.edu/people/c\char"00E9\relaxcile-alduy

[9] Exceptionnal Model Mining, Surpervised descriptive local pattern mining with complex target concepts, Wouter Duivesteijn, Ad J. Feelders et Arno Knobbe, http://hdl.handle.net/1854/LU-7044680

[10] Twython, https://github.com/ryanmcgrath/twython

[11] D3.js, https://d3js.org/

[12] Cortana, Leiden University, http://datamining.liacs.nl/cortana.html

[13] Knime, https://www.knime.org/knime-analytics-platform