

Université de Bourgogne
UFR Sciences et Techniques

MÉMOIRE DE MASTER INFORMATIQUE

BASES DE DONNÉES ET INTELLIGENCE ARTIFICIELLE
2020/2021

par :
Alexis GUYOT

Stage orienté recherche en laboratoire

Polarisation des réseaux sociaux numériques et annotation sémantique transmédia

---

réalisé au sein de l'équipe Sciences des Données du
Laboratoire d'Informatique de Bourgogne (LIB)
9 Avenue Alain Savary 21078 Dijon

sous la direction de :

Éric LECLERCQ
eric.leclercq@u-bourgogne.fr

co-encadré par :

Annabelle GILLET
annabelle.gillet@depinfo.u-bourgogne.fr

enseignant référent :

Nadine CULLOT
cullot.nadine@orange.fr

Remerciements

Tout d'abord, je tiens à remercier mon encadrant, **M. Éric LECLERCQ**, maître de conférences HDR à l'université de Bourgogne, pour m'avoir accompagné, conseillé et guidé avec bienveillance tout au long de ma première année dans le monde de la recherche, du début de mon projet tuteuré à l'obtention de mon financement de thèse I.C.E., en passant bien évidemment par mon stage. Par la même occasion, je remercie grandement ma co-encadrante, **Mme. Annabelle GILLET**, doctorante à l'université de Bourgogne, pour son soutien infaillible, son aide précieuse et surtout pour la transmission passionnée de son expertise sur son langage de programmation de prédilection, Python. L'écoute, l'accompagnement, les explications et la confiance qu'ils m'ont tous les deux accordés m'ont permis d'accomplir en un an plus que je n'aurais pu espérer et d'acquérir une expérience et une vision du milieu et de ses problématiques qui me permettent aujourd'hui d'envisager sereinement mon futur professionnel.

J'adresse également mes remerciements à **M. Olivier TOGNI**, professeur à l'université de Bourgogne et directeur du Laboratoire d'Informatique de Bourgogne (LIB), ainsi qu'à **Mme. Nadine CULLOT**, professeure à l'université de Bourgogne, qui ont tous les deux été particulièrement à mon écoute en cas de besoin et qui me sont venus en aide à plusieurs reprises. De manière plus générale, je remercie toute l'équipe du LIB pour leur accueil très bienveillant.

Ma gratitude est également adressée à **M. Gilles BRACHOTTE**, maître de conférences en sciences de l'information et de la communication et pilote scientifique du projet interdisciplinaire Cocktail, ainsi qu'aux autres membres de l'équipe pour l'expérience que m'a apporté ma participation au processus de recherche ainsi qu'aux diverses réunions de travail et de pilotage du projet.

Enfin, je tiens à exprimer ma reconnaissance à ma famille, à mes amis, à mes anciens professeurs et à toutes ces personnes qui m'ont permis d'en arriver là aujourd'hui, en contribuant de près ou de loin au bon déroulement de mes études.

Alexis Guyot

Table des matières

Table des figures

Chapitre 1

Introduction générale

L'analyse des réseaux sociaux numériques (RSN) fait partie des problématiques actuelles majeures en sciences des données et a pour objectif l'étude de phénomènes, notamment sociologiques, grâce à l'extraction d'informations et de connaissances produites par les utilisateurs de ces applications au travers de la création d'un profil, des interconnexions avec d'autres personnes et de la publication de contenu.

De nombreux projets de recherche traitent aujourd'hui de cette thématique, comme par exemple le projet interdisciplinaire ISITE-BFC Cocktail [1], dans lequel s'inscrit mon stage et qui a pour ambition de mener à la création d'un observatoire en temps réel des tendances, des singularités et des signaux faibles circulant dans les discours de l'alimentaire et de la santé sur les RSN, et plus particulièrement sur Twitter, un réseau social dit de *microblogging* créé en 2006 et qui compte aujourd'hui plus de 199 millions d'utilisateurs quotidiens [2]. La particularité principale de ce RSN réside dans le format des messages textuels pouvant être publiés sur l'application, les *tweets*, dont la taille est limitée à 280 caractères. Cette contrainte force une expression concise et souvent peu nuancée des avis des individus, facilitant l'étude hors contexte de l'environnement social présent sur le RSN.

Le projet est scientifiquement piloté par Gilles Brachotte, maître de conférences en sciences de l'information et de la communication, et réunit principalement des chercheurs en informatique et en sciences humaines et sociales pour l'aspect recherche, et les entreprises dijonnaises AtolCD, Vitagora et Webdrone pour l'aspect industrialisation. Les problématiques de recherche en informatique sont assurées par l'équipe Sciences des Données du Laboratoire d'Informatique de Bourgogne (LIB). Le LIB, dirigé par le professeur des universités Olivier Togni, est une équipe d'accueil de l'université de Bourgogne composée d'une trentaine d'enseignants-chercheurs en informatique et d'une vingtaine de chercheurs contractuels doctorants, ingénieurs et postdoctorants. Le laboratoire est développé autour de trois thématiques principales de recherche : la combinatoire et les réseaux, la modélisation géométrique et la science des données. C'est au sein de cette équipe et sous la supervision d'Éric Leclercq que j'ai effectué mon stage. J'ai été amené à travailler sur deux problématiques liées au projet Cocktail : la détection de polarisation sur Twitter et l'annotation sémantique transmédia.

1. https://projet-cocktail.fr/
2. https://www.journaldunet.com/ebusiness/le-net/1159246-nombre-d-utilisateurs-de-twitter-dans-le-monde/

La polarisation des utilisateurs est un exemple de phénomène pouvant intervenir au fil des discussions sur les RSN et qui peut être identifié par l'étude des discours. Elle intervient lorsque les discussions autour d'un sujet particulier mènent des individus aux opinions similaires à se rassembler au sein de deux groupes principaux qui possèdent des positions opposées, conflictuelles et contrastées, alors que peu d'individus restent neutres ou dans une position intermédiaire. J'ai eu l'occasion de découvrir et de commencer à travailler sur cette thématique dans le cadre de mon projet tuteuré de master, et mes travaux sur le sujet ont mené à la rédaction d'un article en français accepté pour la conférence INFORSID'21 [3]. Les résultats obtenus et les retours très encourageants du comité de lecture de la conférence et des différents acteurs du projet Cocktail nous ont poussé avec Éric Leclercq à décider de consacrer une première partie du stage à prolonger le travail effectué jusque-là. Cette prolongation devait notamment permettre :

— d'étendre ma méthode de détection de polarisation pour la rendre compatible avec les structures communautaires recouvrantes ;

— de mener une nouvelle étude de cas sur des données réelles du projet ;

— de rédiger deux nouveaux articles de recherche, tous les deux en anglais ;

— de préparer et de s'entraîner aux présentations orales des conférences.

L'un des deux nouveaux articles rédigés a déjà pu être soumis et a été accepté par le comité de lecture de la conférence FRCCS [4] 2021. La présentation orale pour cette dernière était préfilmée [5] puis diffusée en différé. Elle devait être faite en anglais avec une durée maximale de 12 minutes. Pour INFORSID'21, la présentation était en direct en visio-conférence, en français et d'une durée de 20 minutes.

J'ai également eu l'occasion de faire partie du comité d'organisation de la conférence FRCCS 2021, qui était organisée par le LIB cette année et qui aurait dû avoir lieu en présentiel à Dijon avant de passer en visio-conférence. Cette expérience fut très enrichissante à trois niveaux, puisqu'il s'agissait de ma première conférence scientifique à la fois en tant que spectateur, orateur et membre du comité local d'organisation. La conférence était entièrement dispensée en anglais, ce qui a ajouté une dimension intéressante à ma participation. En tant que *technical chair*, mes missions au sein de l'organisation étaient les suivantes :

— s'assurer de la qualité audio et du respect des consignes des 22 vidéos contenant les présentations des orateurs qui nous étaient assignés par binôme ;

— être le contact principal des animateurs de sessions (*session chairs*) et des orateurs invités (*keynote speakers*), s'assurer par mail de leur présence, répondre à leurs questions, etc. ;

— manipuler l'outil de visio-conférence Zoom Webinar pour assurer la fluidité de la conférence et assigner correctement les droits lors des 9 sessions qui nous étaient assignées ;

— surveiller le bon déroulement technique, répondre aux questions des spectateurs et assister les responsables de la session lors des passages qui ne nous étaient pas assignés.

Dans un second temps, j'ai pu traiter une nouvelle problématique liée à l'utilisation des données de Twitter, l'enrichissement sémantique, qui consiste à exploiter au mieux les informations à notre disposition pour donner plus de sens aux messages très courts. L'objectif de cette partie du stage était alors d'étudier la possibilité d'annoter sémantiquement à l'aide de mots-clés les

3. INFormatique des ORganisations et Systèmes d'Information et de Décision
4. French Regional Conference on Complex Systems
5. `https://youtu.be/MjWnAMg3PYQ`

médias aux formats hétérogènes (images, vidéos, pages HTML, ...) référencés dans les *tweets*, notamment à partir de méthodes de *machine* ou de *deep learning*. Les objectifs complémentaires à cette thématique étaient :

— d'apprendre à réaliser un état de l'art plus vaste et conséquent que celui rédigé dans le cadre de mon module de master d'initiation à la recherche [6] sur l'anonymisation des graphes [7] ;

— d'appliquer des techniques de *machine learning* ou de *deep learning* ;

— d'exploiter des données volumineuses et de savoir mesurer expérimentalement les performances des solutions proposées.

Pendant mon stage, et plus particulièrement au début de la phase de travail sur l'annotation sémantique, j'ai également été amené à participer à des auditions pour les thèses auxquelles j'avais adressé ma candidature pour pouvoir poursuivre mon cursus universitaire au-delà de mon master. Parmi elle se trouvait la thèse soumise par Éric Leclercq et Nadine Cullot au dispositif Itinéraire Chercheurs Entrepreneurs (ICE) de la région Bourgogne-Franche-Comté, qui finance chaque année un petit nombre (5-7) de thèses pour promouvoir l'émergence d'entreprises à forte valeur ajoutée sur le territoire régional par l'identification et la professionnalisation de futurs chercheurs qui souhaitent suivre un parcours de doctorat intégrant une double compétence recherche et entrepreneuriat/management. Le sujet proposé par Éric Leclercq et Nadine Cullot faisait partie des 11 sélectionnés (sur une cinquantaine) pour les auditions et portait sur la création de lacs de données sémantiques, des systèmes de stockage de masse intelligents faisant usage de méthodes d'apprentissage (*Machine Learning*) et d'intelligence artificielle symbolique pour enrichir et structurer les données qu'ils contiennent avec de nouvelles métadonnées. Quelques jours du stage sur les trois semaines qui ont précédé l'audition face à jury constitué de représentants de la région et du monde professionnel (essentiellement des chefs d'entreprise) ont été dédiés à l'étude approfondie des lacs de données et des métadonnées sémantiques. L'audition a eu lieu le 30 juin dans au siège de l'Université Bourgogne Franche-Comté (UBFC). À son issue, le financement m'a été accordé, ce qui me permettra de continuer en thèse au LIB par la suite, sur un sujet fortement lié aux travaux que j'ai pu effectuer en stage sur l'annotation sémantique.

Enfin, le stage étant orienté recherche, un dernier objectif plus général était la découverte du milieu et du métier d'enseignant-chercheur, par exemple en participant à la vie du laboratoire. À ce niveau-là, j'ai eu l'occasion de participer et de présenter mon travail lors d'une réunion de travail ainsi que lors d'une réunion de pilotage du projet Cocktail, en compagnie des collègues chercheurs en sciences humaines et sociales pour les deux et des représentants des acteurs industriels lors de la seconde. J'ai également eu la possibilité de parler de mes travaux de recherche lors d'un séminaire du LIB de 45 minutes.

Une chronologie permettant d'observer la répartition du temps de travail par partie du stage est proposée sur la figure 1.1. Pendant la phase "Conférences", le temps de travail non dédié à la préparation des présentations ou à l'organisation de FRCSS 2021 était globalement plutôt centré sur la polarisation et la rédaction des articles.

Dans la suite de ce document, je présente les résultats des travaux effectués dans le cadre de mon stage. Dans un premier temps, le chapitre 2 détaille à la fois ma contribution à la détection de polarisation sur Twitter par l'étude des frontières de communautés d'utilisateurs, mise à jour

6. Module d'enseignement du parcours recherche qui propose d'étudier un corpus d'une dizaine d'articles choisis par les encadrants sur une thématique donnée puis de restituer les connaissances acquises à la fois à l'écrit via la rédaction d'un état de l'art et à l'oral via une présentation.

7. `https://github.com/AlexisGuyot/anonymisation-graphes`

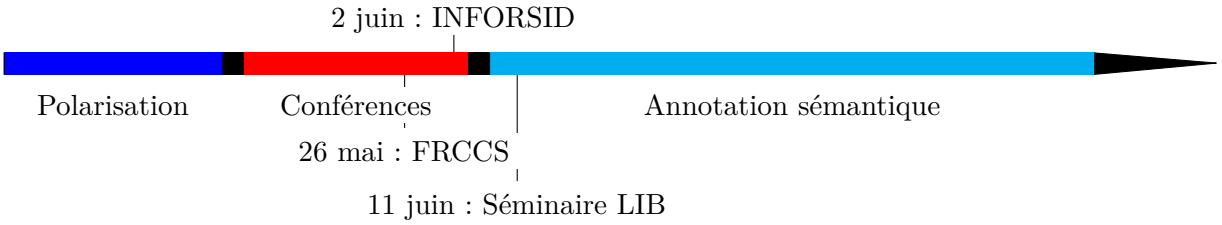


FIGURE 1.1 – Chronologie du stage d'avril à septembre 2021

avec les extensions travaillées lors du stage, et les résultats obtenus lors de la nouvelle étude de cas. Dans un deuxième temps, je dresse dans le chapitre 3 un état de l'art de l'extraction automatique de mots-clés, technique utilisée pour annoter sémantiquement les données non-structurées textuelles. Dans cette même partie, je présente quelques expérimentations mises en place à partir des connaissances acquises dans la littérature et ayant pour but d'identifier l'approche de l'état de l'art la plus adaptée à un jeu de données du projet Cocktail. Enfin, je présente dans le chapitre 4 une conclusion générale aux travaux effectués lors de mon stage, en discutant de l'expérience que j'ai pu acquérir et des perspectives de ce travail, notamment en faisant le lien avec ma future thèse que je commencerai début octobre au LIB.

Chapitre 2

Frontières des communautés d'individus et polarisation des réseaux sociaux numériques

2.1 Introduction

Depuis sa création au début des années 90, le *World Wide Web* n'a cessé d'évoluer et de s'adapter à ses utilisateurs. Ces derniers, à l'origine seulement consommateurs de contenu, sont, au fur et à mesure du temps, devenus producteurs avec l'arrivée du Web 2.0 [116], aussi connu sous le nom de Web Participatif. Un aspect important de cette évolution est l'émergence du Web Social [130], où l'information sociale créée par les personnes et leurs interactions devient le centre de l'attention. Sont alors apparus les réseaux sociaux numériques (RSN), des plateformes où les utilisateurs peuvent partager des informations personnelles via la création de profils et interagir entre eux par différents moyens.

En 2021, 50.64% [1] de la population mondiale utilise activement au moins un RSN. Cette statistique atteint même 83.36% si on ne considère que l'audience éligible, c'est-à-dire les 4,59 milliards d'utilisateurs du Web. En capturant une très grande quantité d'interactions entre les individus, les RSN forment des environnements sociaux qui peuvent ensuite être étudiés par le biais d'approches dites de SNA pour *Social Network Analysis* (en français littéralement l'Analyse des Réseaux Sociaux). Ces études trouvent de nombreuses applications dans des domaines divers et variés comme la sociologie, le marketing ou encore la politique.

Les méthodes de SNA exploitent pour la majeure partie les connaissances et algorithmes de la théorie des graphes [156]. Par conséquent, l'information sociale extraite des RSN est représentée à l'intérieur de structures de ce type. Les utilisateurs sont vus comme des sommets et les interactions entre eux comme des liens dirigés et pondérés. L'exploitation des graphes permet ensuite aux analystes de mieux comprendre l'environnement social, par exemple en étudiant le rôle des différents utilisateurs grâce aux mesures de centralités [21], ou encore en s'intéressant aux phénomènes de diffusion de l'information.

La plupart des graphes extraits des RSN, aussi appelés graphes sociaux, possèdent une structure particulière caractérisée par une distribution des degrés de leurs sommets (nombre de liens entrants et sortants) qui suit une loi de puissance. Cette propriété fait d'eux des réseaux sans-échelle [9] et rend possible la détection et l'étude de communautés de sommets [49, 89, 129], et

1. `https://backlinko.com/social-media-users`

donc par extension d'individus, afin de mieux comprendre les différents rôles et comportements au sein du RSN. Ces communautés correspondent à des zones localement denses à l'intérieur des graphes, où le nombre de liens entre les sommets d'une même communauté est élevé par rapport au nombre de liens qui en sortent.

La détection de communautés est un problème bien connu en sciences des données, avec de très nombreuses applications dans des domaines divers et variés : analyse des RSN [18, 35, 125, 127, 133], des réseaux du monde réel [159], d'Internet [44], de ceux du Web [85] et même de certains réseaux biologiques [52]. La plupart du temps, les communautés détectées sont utilisées telles quelles et des mesures comme la modularité [113], qui sert notamment à évaluer la cohésion de la structure communautaire, sont calculées directement sur elles. Cependant, pour une meilleure interprétation de certains phénomènes, une vue plus détaillée des groupes d'individus est parfois nécessaire.

Dans la littérature, les frontières des communautés sont des zones très peu explorées. On peut grossièrement définir une frontière comme l'ensemble des membres d'une communauté qui possèdent au moins un lien avec un sommet qui appartient à leur communauté et un autre avec un sommet qui n'appartient pas à la communauté. Cette définition a notamment été utilisée par Clauset [34] dans sa méthode d'extraction locale de communautés. L'analyse des frontières est une problématique très prometteuse en SNA. En effet, les individus qui peuplent les frontières déterminent l'ouverture de la communauté aux autres et révèlent ainsi sa place et son rôle au sein du réseau. Elle permet également de mieux comprendre comment l'opinion du groupe est partagée et défendue face à d'autres points de vue. Selon le type d'interaction représentée dans le graphe, des traces d'antagonisme peuvent même être détectées lors des échanges, comme exploré par Guerra et al. [58] sur des graphes sociaux non-dirigés et non-pondérés.

Avec des informations relatives à la cohésion et au degré d'opposition entre groupes, il devient possible de détecter des traces de polarisation entre paires de communautés. En sciences sociales, ce phénomène a été défini par Isenberg comme le processus qui mène un groupe social à se diviser en deux sous-groupes opposés possédant des opinions conflictuelles et contrastées sur un sujet donné, avec peu d'individus qui restent neutres ou dans une position intermédiaire [73, 143]. Détecter les sujets qui polarisent les RSN est considéré aujourd'hui comme une des clés principales pour mieux appréhender et comprendre l'opinion publique, afin de pouvoir correctement adapter en conséquence des stratégies de communication ou des décisions marketing, ou même combattre la désinformation.

Pendant mon stage, et avant cela pendant mon projet tuteuré, j'ai eu l'occasion de travailler sur et d'approfondir l'utilisation des frontières pour l'étude des comportements des communautés au sein des graphes sociaux, en me basant sur les travaux initiés en 2013 par Guerra et al. [58]. Dans le but d'obtenir des informations plus précises et de proposer une méthode plus proche de la réalité de terrain induite par les données réelles, j'ai ainsi travaillé lors de mon projet tuteuré sur une extension de ces travaux en prenant en considération la direction et la pondération des graphes. Dans le cadre de mon stage, j'ai décidé d'aller encore plus loin en rendant ma méthode compatible avec les structures communautaires dites recouvrantes, où un même sommet peut appartenir à une ou plusieurs communautés en même temps. Une implémentation en R de ma méthode est accessible en libre accès sur mon GitHub [2].

Mes travaux sur la thématique de ce chapitre ont mené à la rédaction d'un article long en français pendant mon projet tuteuré, qui a été accepté pour publication par le comité de lecture

2. `https://github.com/AlexisGuyot/CommunityBoundaries`

de la conférence INFORSID 2021. Pendant mon stage, j'ai eu l'occasion d'écrire un nouvel article, court et en anglais, qui a aussi été accepté pour publication par la conférence FRCCS 2021. J'ai été amené à présenter oralement mes travaux lors de ces deux conférences, qui ont toutes les deux eu lieu pendant mon stage, respectivement début juin et fin mai. Nous avons également travaillé avec Éric Leclercq et Annabelle Gillet sur la rédaction d'un troisième article long et en anglais à destination d'une conférence internationale, qui sera soumis après la fin de mon stage.

Dans la section 2.2, un rapide état de l'art sur la détection de polarisation sur les RSN est dressé par la présentation de quelques travaux connexes. Ensuite, après quelques rappels des concepts fondamentaux de la théorie des graphes dans la section 2.3, je détaille ma méthode dans la section 2.4. Dans la section 2.5, je présente les résultats obtenus lors d'une étude de cas menée dans le cadre du projet interdisciplinaire Cocktail et qui m'a permis d'utiliser ma méthode pour commenter une possible polarisation de Twitter autour de la question de la vaccination contre la Covid-19. Enfin, je conclus ce chapitre sur la polarisation dans la section 2.6 en ouvrant sur quelques perspectives de travail issues de pistes envisagées ou explorées sans aboutir à de réels résultats lors de mon stage.

2.2 Travaux connexes

La polarisation des communautés est une thématique très importante en SNA depuis plusieurs années maintenant. On retrouve ainsi des travaux sur le sujet qui remontent à 2011 avec par exemple Conover et al. [37] sur Twitter. Plusieurs autres approches ont depuis été proposées pour détecter des traces de polarisation, en utilisant à chaque fois différentes caractéristiques et/ou différents éléments du contenu ou de la structure des graphes sociaux.

2.2.1 Approches avec traitement du langage naturel

Une analyse hybride de la structure du graphe et des sentiments exprimés dans les messages est proposée dans les articles de Alamsyah et Adityawarman [3] et de Habibi et al. [61]. La méthodologie des deux articles consiste à exécuter en parallèle deux processus :

— une analyse de la structure du graphe, composée d'une détection de communautés et du calcul de mesures de centralité ;

— une classification des individus par un algorithme de Traitement du langage naturel (TLN), selon leur propension à publier des *tweets* majoritairement catégorisés comme pro, anti ou neutres par rapport à la thématique étudiée.

Les résultats des deux processus sont analysés conjointement, notamment en vérifiant la composition des différentes communautés grâce à la classification basée sentiments. Les deux méthodes utilisent une classification naïve bayésienne et atteignent des scores finaux de précision de 97.42% pour la première et entre 69.2 et 100% pour la deuxième.

Même si les valeurs précédentes sont assez élevées, Habibi et al. [61] les nuancent en soulignant l'importance de la taille du jeu de données et de l'étape de nettoyage qui précède la classification. À cause de cette contrainte, leur méthodologie ne peut pas être utilisée dans toutes les circonstances ou de manière complètement automatique. En effet, le nettoyage des données est une étape difficile qui nécessite en général l'investissement d'un être humain pour gérer le grand nombre de difficultés et ambiguïtés du langage naturel des messages issus des RSN comme les approximations orthographiques, abréviations, argot, etc. L'aspect hors-contexte peut

également rendre difficile pour une machine la détection d'un objectif humoristique ou la présence de sarcasme ou d'ironie, comme le montrent les travaux de González-Ibánez, Muresan et Wacholder [55], Joshi, Bhattacharyya et Carman [82] et McGlone [103]. De plus, puisque différentes langues cohabitent sur les RSN, et même parfois au sein d'un même *tweet*, la barrière de la langue peut aussi être un problème avec ce type de méthodes.

Malgré tout, l'utilisation du TLN en complément des analyses structurelles alimentées par la théorie des graphes reste assez courante. C'est par exemple le cas avec la méthode de Jiang, Ren et Ferrara [76], qui utilise *Retweet-BERT*, un modèle de plongement de phrases assez réputé dans la littérature, couplé avec une étude de voisinage, dans le but d'attribuer une valeur de polarisation à chaque sommet d'un réseau de *retweets* et d'un réseau de citations. Ces valeurs sont ensuite utilisées pour mener des études précises sur les différents individus des réseaux en fonction de leur rôle (partisans ou influenceurs). Les auteurs parviennent ainsi à analyser le comportement des membres de leurs différents pôles et à pointer leur propension à devenir des chambres d'écho.

2.2.2 Approches faiblement ou non-supervisées

Pour éviter les difficultés et restrictions induites par l'utilisation du TLN, des stratégies plus faiblement supervisées sont également envisagées. L'intervention de l'analyste lors du processus n'est alors plus déterminante à son exécution mais seulement à l'obtention de résultats pertinents.

Concrètement, cela peut se résumer à correctement initialiser un algorithme, comme le proposent Morales et al. [109]. Pour utiliser leur méthode, l'analyste doit en amont manuellement attribuer un label égal à $+1$, 0 ou -1 aux n individus les plus centraux du graphe (appelés *elites*) afin de décrire s'ils sont respectivement pro, neutres ou opposés au sujet dont on veut vérifier le degré de polarisation. Les valeurs attribuées sont ensuite utilisées pour calculer celles des autres sommets du graphe social (appelés *listeners*) par propagation. Chaque valeur obtenue positionne l'individu sur un spectre d'opinion et décide de son appartenance à un groupe ou à un autre : les sommets avec une valeur comprise entre α et $-\alpha$, avec α une marge choisie par l'analyste et comprise entre 0 et 1, sont catégorisés comme neutres, ceux avec une valeur supérieure à α comme pro, ceux avec une valeur inférieure à $-\alpha$ comme anti. Une mesure de polarisation pour le graphe est ensuite calculée en prenant en considération la taille des deux groupes et la distance entre les centres de gravité (approche inspirée de la notion de polarisation électrique). La phase d'initialisation de cette méthode constitue sa plus grande limite, puisque la pertinence de ses résultats reste très sensible aux valeurs attribuées aux sommets *elites*.

Une approche non-supervisée d'extraction des groupes polarisés d'un graphe social à l'aide d'une factorisation de matrices et d'un algorithme de descente de gradient ensembliste est explorée par Al Amin et al. [2]. La méthode utilise un graphe biparti qualifié de graphe source-assertion. Les sources sont des individus du RSN et les assertions des clusters de *tweets* similaires. Est également utilisé un graphe social d'influence qui représente la tendance de chaque source à partager une assertion. Les deux structures de données précédentes sont utilisées conjointement pour calculer deux matrices U et V contenant respectivement les probabilités pour chaque source et pour chaque assertion d'appartenir aux différents groupes polarisés. Les auteurs montrent dans leur article que leur algorithme réussit à bien mieux extraire les groupes polarisés que les méthodes basées sur une analyse de sentiments. Ils montrent aussi que leur méthode permet d'obtenir des résultats 20 à 30% meilleurs que les approches basées sur une simple détection de communautés quand le graphe contient beaucoup de sources et d'assertions neutres. La princi-

pale limite de cette méthode réside dans le choix de la mesure de similarité pour construire les assertions. En effet, la meilleure alternative dépend du jeu de données et des connaissances a priori sur les données sont donc nécessaires pour départager. La méthode ne donne également aucune indication concernant le degré d'opposition ou d'antagonisme exprimé entre les groupes polarisés. Pourtant, cette connaissance est très importante car elle permet de distinguer des communautés polarisées de simples chambres d'écho s'ignorant ou n'ayant pas connaissance des unes et des autres.

2.2.3 Approches structurelles

En 2013, Guerra et al. [58] proposent d'étudier les frontières des communautés pour obtenir des informations sur la polarisation. Dans leur méthode, un sommet est considéré comme frontière de sa communauté $C1$ avec la communauté $C2$ si au moins un de ses voisins appartient à $C2$ et qu'au moins un autre appartient lui aussi à $C1$ mais sans posséder de voisin dans $C2$. Ils émettent l'hypothèse que plus un individu est impliqué au sein de sa communauté, plus l'opinion du groupe est susceptible de lui tenir à cœur, ce qui lui donne plus de chances d'être vecteur d'antagonisme lors de ses interactions avec les membres d'autres communautés. En partant de ce principe, leur méthode permet de calculer une valeur d'antagonisme pour chaque sommet frontière grâce à une formule dérivée de la modularité locale proposée par Clauset [34], c'est-à-dire une proportion entre le nombre d'interactions internes et le nombre total d'interactions internes et externes en provenance du sommet. Le score final pour une paire de communauté donnée est obtenu en calculant la moyenne des scores d'antagonisme des différents membres de la zone frontière. Les auteurs préconisent d'utiliser cette valeur en complément de la modularité, mesure proposée par Newman [113] qui permet d'évaluer la cohésion des différentes communautés en étudiant leur densité par rapport à un sous-graphe de même taille dans un réseau aléatoire. La méthode de Guerra et al. [58] est formalisée pour des graphes non-dirigés et non-pondérés, ce qui rend leur notion d'antagonisme symétrique. Or, ce fait est réfutable car un individu peut simplement ignorer son opposant plutôt que de répondre. Enfin, la méthode ne prend en charge que le cas des communautés disjointes, où chaque sommet n'appartient qu'à une et à une seule communauté. Cependant, certaines études montrent qu'en réalité, les utilisateurs des RSN appartiennent en général à plusieurs communautés en même temps [161].

2.2.4 Synthèse

En prenant en considération les différents travaux de recherche réalisés sur la polarisation des communautés sur les RSN, j'ai décidé de m'orienter vers les approches uniquement structurelles basées sur l'analyse des frontières des communautés pour plusieurs raisons :

— proposer une méthode **applicable** dans tous ou presque **tous les cas**, même sans connaissance a priori sur le graphe social ;

— permettre l'intégration de la méthode au sein de **processus automatiques d'analyse** plus complexes (par exemple des logiciels ou applications) ;

— faciliter la **reproductibilité** des résultats obtenus.

Toutes ces raisons font également écho aux besoins et contraintes liées au projet Cocktail et au bon développement d'un observatoire en temps réel.

Les méthodes basées sur **l'analyse des frontières des communautés** permettent de commenter de manière assez précise la polarisation d'un graphe social en fournissant des informations à la fois sur la **cohésion des groupes** et sur la **nature de leurs interactions**, ce qui couvre les principaux facteurs de polarisation énoncés par Isenberg [73]. Cependant, les travaux de Guerra et al. [58] ne se sont concentrés jusque-là que sur les graphes non-dirigés et non-pondérés, alors que les interactions entre individus sur les RSN sont en général plutôt **dirigées**, avec un émetteur et un destinataire, et **pondérées** avec le nombre d'interactions. De plus, le cas des **structures communautaires** dites **recouvrantes**, où les sommets du graphe social peuvent potentiellement appartenir à plusieurs communautés simultanément n'a, au vu des méthodes présentées dans cette section et de mes recherches personnelles, jamais été traité conjointement à la recherche de polarisation dans la littérature, malgré les études montrant la pertinence de leur association. Ma contribution, développée à la fois à travers mon projet tuteuré puis mon stage, consiste donc à surpasser ces limitations en proposant une **nouvelle approche générique et non-supervisée** inspirée des travaux de Guerra et al. [58], mais **plus adaptée aux données réelles** issues des réseaux sociaux, comme celles collectées dans le cadre du projet Cocktail.

2.3 Rappel des concepts fondamentaux

Un **graphe** G = (V,E) est composé d'un ensemble de composants V de taille $|V|$, appelés sommets ou nœuds, et d'un ensemble d'interactions ou de connexions E de taille $|E|$ qui les relient entre eux, appelés arêtes ou liens. Une arête e_{ab} entre un sommet a et son voisin b peut être dirigée, a est alors qualifié de source et b de destination, ou non-dirigée et permettre la navigation dans les deux sens. Une arête peut également posséder un poids $w(e_{ab})$ pour évaluer la force de la connexion. La structure d'un graphe peut être représentée à l'intérieur d'une liste d'adjacence Al, un dictionnaire qui associe à chaque sommet sa liste de voisins, ou à l'intérieur d'une matrice d'adjacence Am, une matrice carrée dans laquelle chaque cellule contient une valeur binaire marquant l'existence d'un lien dont la source est le sommet qui indexe la ligne et la destination celui qui indexe la colonne. Dans le cas d'un graphe pondéré, les valeurs contenues dans les cellules ne sont pas binaires mais décimales et représentent le poids du lien.

Une **communauté** C_i est un sous-ensemble de sommets localement dense : le nombre de liens reliant les sommets du sous-ensemble est élevé par rapport au nombre de liens en direction de sommets situés à l'extérieur du sous-ensemble. Un graphe contient $|C|$ communautés $C = \{i, j, ..., n\}$. Si chaque sommet du graphe appartient à une et une seule communauté, c'est-à-dire que $\forall (C_a, C_b) \in C^2 \mid C_a \cap C_b = \emptyset$, alors on parle de structure communautaire disjointe (ou plus simplement de communautés disjointes). Autrement, si chaque sommet peut appartenir à une ou plusieurs communautés simultanément, on parle de communautés recouvrantes.

2.4 Présentation de la méthode

Dans cette section, la méthode que je propose pour évaluer le degré de polarisation des communautés des graphes sociaux en étudiant leurs frontières est détaillée par la description des différents ensembles et formules utilisés, de quelques approches algorithmiques possibles et par une comparaison des résultats obtenus avec ma spécification et de ceux obtenus avec l'approche de Guerra et al. [58], dans le but de justifier l'intérêt de mon extension sur un exemple simple.

2.4.1 Ensembles et formules

Dans un premier temps, la méthode consiste à analyser le voisinage de chaque membre des communautés afin de toutes les décomposer en deux zones d'intérêt. Pour chaque paire de communautés $(C_i, C_j) \in C^2$, on identifie :

— la zone interne $I_{i,j}$ de C_i, qui correspond à l'ensemble des sommets de la communauté qui ne possèdent pas de voisin membre de C_j ;

— la zone frontière $B_{i,j}$ de C_i, qui correspond à l'ensemble des sommets de la communauté qui possèdent au minimum un voisin membre de C_j et un voisin membre de $I_{i,j}$.

Ces définitions se rapprochent de celles proposées par Guerra et al. [58]. J'ai veillé toutefois à corriger une inconsistance entre les ensembles proposés par les auteurs et la description textuelle et via l'exemple qu'ils en font dans leur article. Comme dans la méthode précédente, une restriction supplémentaire est ajoutée sur les sommets frontières : ils doivent posséder au moins un voisin de chaque communauté de la paire étudiée (comme dans l'approche de Clauset [34]), et en plus de cela le voisin de la même communauté doit faire partie de la zone interne et donc ne pas avoir de connexion à l'autre communauté. Cette restriction agit en réalité comme un filtre adapté aux interactions sur Twitter. En effet, les publications sur ce RSN étant publiques et facilement accessibles par n'importe qui, il est possible qu'un utilisateur se retrouve par erreur associé à une communauté dont il ne fait pas partie. Par exemple, s'il prend part à une discussion avec un individu frontière ou s'il partage du contenu produit par la communauté à des fins d'humour ou de sarcasme sans être particulièrement associé à un autre groupe d'utilisateurs. En s'assurant que chaque individu frontière interagisse à la fois avec l'extérieur et avec des membres très renfermés de leur communauté, les individus "intrus" sont ignorés et leur présence n'engendre pas de bruit dans les résultats.

En prenant en considération les éléments précédents, je propose deux nouvelles définitions formelles des zones internes et frontières dans les équations 2.1 et 2.2.

$$I_{i,j} = \{v : v \in C_i, \nexists e_{vn} \mid n \in C_j, i \neq j\} \tag{2.1}$$

$$B_{i,j} = \{v : v \in C_i, \exists e_{vn_1} \mid n_1 \in C_j, \exists e_{vn_2} \mid n_2 \in I_{i,j}, i \neq j\} \tag{2.2}$$

Les concepts de liens faibles et de liens forts de la littérature [56, 99, 157] sont utilisés pour calculer la valeur d'antagonisme. À l'origine, les liens faibles relient deux sommets situés dans des communautés différentes, là où les liens forts relient deux sommets d'une même communauté. Pour ma méthode, je propose de rendre ces définitions plus spécifiques pour distinguer les arêtes qui sortent d'une frontière. Ainsi, les liens faibles relient un sommet frontière et un sommet de l'autre communauté (E_B), là où les liens forts relient un sommet frontière et un sommet interne (E_{int}). Les nouvelles définitions formelles que je propose pour les liens faibles et forts sont présentées dans les équations 2.3 et 2.4.

$$E_B = \{e_{sd} : s \in B_{i,j} \wedge d \in C_j\} \tag{2.3}$$

$$E_{int} = \{e_{sd} : s \in B_{i,j} \wedge d \in I_{i,j}\} \tag{2.4}$$

Pour mieux appréhender les différents ensembles, la figure 2.1 présente un graphe exemple dirigé et pondéré contenant trois communautés recouvrantes C = {Bleu, Rouge, Violet}. Les zones internes et frontières ainsi que les liens faibles et forts sont identifiés.

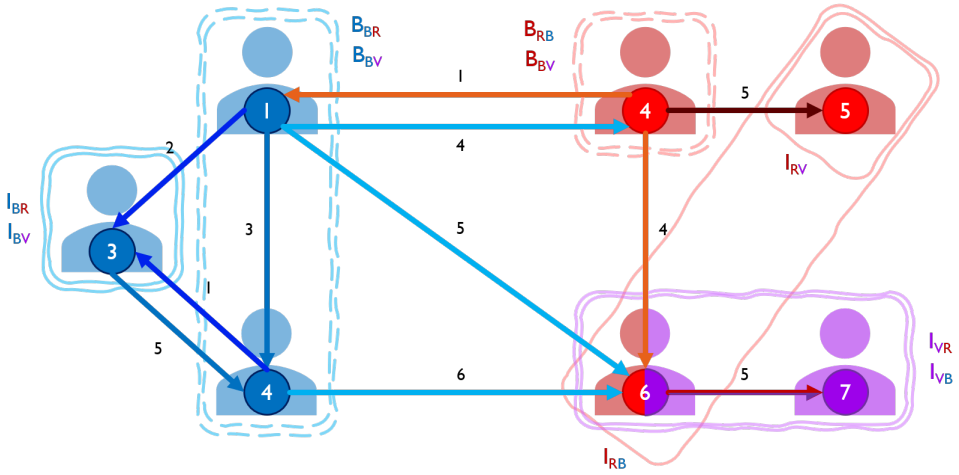


FIGURE 2.1 – Graphe exemple - Les zones internes sont entourées par des lignes solides, les zones frontières par des lignes pointillées. Les liens faibles sont plus clairs et les liens forts plus foncés (meilleur rendu en couleurs).

Le score d'antagonisme pour un sommet frontière donné est le ratio pondéré du nombre de liens forts et de la somme du nombre de liens faibles et forts dont le sommet est source. Le nombre de liens est à chaque fois pondéré par le poids de l'arête. Dans le but de retirer le bruit aléatoire présent dans les graphes sociaux, ce ratio est ensuite comparé par soustraction à l'hypothèse nulle suivante : chaque sommet interagit autant avec sa zone interne qu'avec l'autre communauté (identique à GUERRA et al. [58] [58]). Pour une frontière de communauté donnée, son score d'antagonisme S correspond à la moyenne des scores d'antagonisme de ses sommets membres, comme présenté dans l'équation 2.5.

$$S = \frac{1}{|B_{i,j}|} \sum_{v \in B_{i,j}} \left[\frac{\sum_{e \in E_{iv}} w(e)}{\sum_{e \in E_{iv}} w(e) + \sum_{e \in E_{bv}} w(e)} - 0.5 \right] \tag{2.5}$$

$$E_{iv} = \{e_{vd} : e_{vd} \in E_{int}\} \tag{2.6}$$

$$E_{bv} = \{e_{vd} : e_{vd} \in E_B\} \tag{2.7}$$

Le score d'antagonisme S est compris dans l'intervalle [-0.5, 0.5]. Une valeur positive indique que la frontière interagit plus avec l'intérieur de sa communauté qu'avec l'extérieur, et inversement. Selon l'hypothèse formulée et vérifiée dans l'article de GUERRA et al. [58] concernant l'antagonisme et l'investissement des frontières au sein de leur communauté, un sommet ou une frontière avec un score d'antagonisme positif est plus susceptible d'être vecteur d'antagonisme et donc de contribuer à la polarisation des communautés C_i et C_j. Dans l'utilisation qu'ils font de leurs résultats, GUERRA et al. [58] vont même plus loin en assurant qu'une valeur négative est signe d'une absence de polarisation. Cependant, certains travaux sur le sujet montrent que dans certains cas, une exposition forte aux opinions divergentes d'autres personnes peut aussi mener à une polarisation de la communauté [8, 83, 91], ce qui invaliderait cette dernière affirmation.

En mesurant les scores d'antagonisme des frontières pour chaque paire de communautés, on obtient une matrice carrée de taille $|C| \times |C|$ appelée matrice d'antagonisme. Puisque cette approche prend en considération la direction des liens, cette matrice est asymétrique. Ainsi, les

paires de communautés (C_i,C_j) et (C_j,C_i) doivent toutes les deux être prises en considération et leur score calculé. La figure 2.2 montre la matrice d'antagonisme obtenue par analyse du graphe exemple de la figure 2.1.

	Blue	Red	Violet
Blue	0,000	-0,338	-0,286
Red	0,400	0,000	0,055
Violet	0,000	0,000	0,000

FIGURE 2.2 – Matrice d'antagonisme du graphe exemple.

Pour finir, la détection des zones et types de liens permet également de calculer un autre indicateur appelé porosité des frontières. Dans une frontière $B_{i,j}$ construite à partir de la paire de communautés (C_i,C_j), si on considère B_p le sous-ensemble de sommets de $B_{i,j}$ qui possèdent un score d'antagonisme positif et B_n le sous-ensemble de sommets qui possèdent un score négatif ou nul, la porosité $P_{i,j}$ de la frontière est donnée par l'équation 2.8.

$$P_{i,j} = \frac{B_n}{B_p + B_n} \times 100 \tag{2.8}$$

2.4.2 Algorithme

Afin de pouvoir appliquer ma contribution sur des données réelles pour des études de cas liées au projet Cocktail, réfléchir à puis développer un algorithme était une nécessité. Cette partie peut également être considérée comme une contribution supplémentaire, dans le sens où l'article de GUERRA et al. [58] ne propose pas d'algorithme pour expérimenter leur proposition.

Une approche naïve pour calculer la matrice d'antagonisme pourrait être de prendre en considération chaque paire de communautés du graphe une par une et de remplir à chaque fois des ensembles de sommets en analysant la liste d'adjacence, possiblement récursivement pour vérifier que les voisins des sommets frontières situés dans la zone interne ne possèdent pas de voisin de l'autre communauté. Cependant, cette approche implique une très grande complexité algorithmique, ce qui peut être une contrainte sur des graphes de grande taille.

Pour réduire la complexité, la solution que j'ai développée utilise une structure intermédiaire qui décrit les rôles joués par chaque sommet (interne ou frontière) au sein de sa ou de ses communautés par rapport à chaque autre avec un système de code. Cette matrice structurelle est formée de $|V|$ lignes et de $|C|$ colonnes et doit être remplie avant le calcul de la matrice d'antagonisme. Pour plus de détails sur l'algorithme complet, son implémentation en R est disponible en libre accès sur mon GitHub [3].

2.4.3 Différences avec Guerra et al.

Pour rappel, la méthode présentée ici peut être considérée comme une extension des travaux de GUERRA et al. [58] qui a pour but d'améliorer l'adéquation de l'approche avec les données

3. `https://github.com/AlexisGuyot/CommunityBoundaries`

réelles. Pour cela, la direction et la pondération des liens sont prises en considération lors du calcul des scores d'antagonisme, tout comme la possibilité de travailler avec des communautés recouvrantes. Dans cette sous-section, je vais expliquer ce qu'apportent ces éléments à l'approche de Guerra et al. [58] à l'aide de l'exemple simple de la figure 2.1.

D'abord, pour pouvoir appliquer la méthode de Guerra et al. [58], la structure doit être transformée pour obtenir un graphe non-orienté, non-pondéré et contenant uniquement des communautés disjointes. Un résultat possible de cette transformation est présenté dans la figure 2.3.

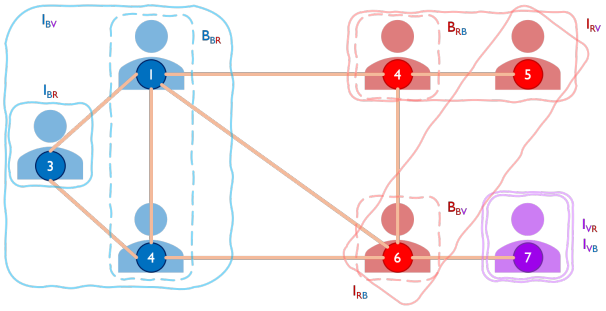


Figure 2.3 – Le graphe exemple transformé pour correspondre à la réalité étudiée par la méthode de Guerra et al. [58] sans direction des liens, pondération et communautés recouvrantes.

	Blue	Red	Violet
Blue	0,000	0,222	0,000
Red	0,222	0,000	0,000
Violet	0,000	0,000	0,000

(a) résultats du graphe transformé (Guerra et al. [58])

	Blue	Red	Violet
Blue	0,000	-0,338	-0,286
Red	0,400	0,000	0,055
Violet	0,000	0,000	0,000

(b) résultats du graphe exemple non-transformé

Figure 2.4 – Matrices d'antagonisme des deux graphes.

Les résultats présentés sur la figure 2.4 illustrent l'impact de la transformation. Par rapport à la matrice 2.4b, on remarque sur la matrice 2.4a l'absence complète d'informations d'antagonisme concernant la communauté violette. La disparition de la structure communautaire recouvrante est principalement à l'origine de cette observation. Lors de la transformation, seule une des deux communautés originellement attribuées au sommet 6 est restée, ici en l'occurrence la communauté rouge, ce qui a eu pour conséquence de retirer à la communauté violette ses connexions avec la communauté bleue. Les définitions ensemblistes de Guerra et al. [58], qui stipulent que seuls les liens faibles entre deux frontières doivent être considérés, sont à l'origine de l'absence de valeurs avec la communauté rouge. En effet, puisque la communauté violette ne possède pas de zone frontière, son lien avec la communauté rouge ne compte pas.

Sur la matrice 2.4a, la relation entre la communauté rouge et la communauté bleue est symétrique à cause de l'absence de direction des liens, qui ne permet pas de savoir si les interactions vont plus dans un sens que dans un autre. Or, on peut voir sur la figure 2.1 que les deux communautés n'interagissent pas entre elles de la même manière : les individus frontières de la communauté bleue interagissent beaucoup avec des membres de la communauté rouge, mais pas

l'inverse. Selon la théorie faite par Guerra et al. [58] concernant l'utilisation de l'investissement des membres frontières pour évaluer l'antagonisme entre deux communautés, seule la communauté rouge est très susceptible d'être antagoniste. Cependant, seule la prise en considération de la direction des liens permet de remarquer cette nuance. Cette simple différence a pourtant un impact important sur l'interprétation des résultats. En effet, la matrice 2.4a montre un antagonisme mutuel entre les communautés bleue et rouge. En prenant la définition de Isenberg [73] comme référence, on conclura alors que ces deux communautés sont polarisées après application de la méthode de Guerra et al. [58], mais pas avec l'extension présentée dans ce document.

Enfin, la dernière différence notable entre les deux matrices concerne les valeurs d'antagonisme elles-mêmes. On note des valeurs très proches des bornes sur la matrice 2.4b, mais pas sur l'autre. Si on se réfère au graphe exemple, les poids des arêtes nous apprennent qu'il y a en réalité peu de liens forts entre les sommets bleus (1 entre les sommets 4 et 3 et 2 entre les sommets 1 et 3) et un grand nombre de liens faibles avec la communauté rouge (4 entre les sommets 1 et 4, 5 entre 1 et 6 et 6 entre 4 et 6). Inversement, la communauté rouge possède beaucoup de liens forts et peu de liens faibles avec la communauté bleue. Sur le graphe transformé, un lien est présent entre deux sommets s'ils ont déjà interagi entre eux, peu importe le nombre de fois. L'élément le plus important pour déterminer la valeur d'antagonisme n'est alors plus le nombre d'interactions, mais le nombre d'individus avec qui le sommet frontière interagit. Contrairement aux autres différences, il est ici plus difficile de savoir si la simplification exigée par la méthode de Guerra et al. [58] implique une perte d'information. La formule pour calculer l'antagonisme est basée sur une évaluation de l'investissement, ce qui signifie que pour obtenir une réponse il faut avant tout répondre à une autre question : un individu est-il plus impliqué s'il interagit un grand nombre de fois avec peu de personnes ou avec un grand nombre de personnes en ayant possiblement peu d'interactions avec chacune ? Les deux approches défendent une option différente, mais la réalité doit plutôt se situer quelque part entre les deux. Une prolongation possible de ce travail pourrait consister à trouver une manière de modifier la formule d'antagonisme pour mieux capturer la notion d'investissement, qui pourrait par exemple prendre en considération à la fois le nombre d'interactions via la pondération et le nombre d'individus différents concernés.

2.5 Nouvelle étude de cas

Les résultats obtenus lors de la dernière étude de cas que j'ai décidé de mener dans le cadre du projet Cocktail sont détaillés dans cette section. Ceux-ci ont notamment pu être présentés aux collègues de Sciences Humaines et Sociales lors d'une réunion.

2.5.1 Contexte et jeu de données

Dans le contexte du projet Cocktail, un jeu de données de plus de 18 millions de *tweets* en français contenant un ou plusieurs mots-clés liés à la Covid-19 et aux vaccins a été collecté grâce à l'architecture Hydre [51] entre le 1^{er} décembre 2020 et le 31 mars 2021 (120 jours). Ces *tweets* peuvent prendre la forme de *tweets* originaux, de *retweets* (partages), de *quotes*/citations (partages avec commentaire supplémentaire) ou de réponses.

À partir de ce jeu de données, j'ai extrait le graphe des citations $G_Q = (V, E)$. Il s'agit d'un graphe dirigé et pondéré dans lequel chaque sommet $u \in V$ est un utilisateur de Twitter et chaque arête $(u, n) \in E$ de poids $w \in \mathbb{N}^*$ représente le fait que l'utilisateur u a cité w fois un ou plusieurs *tweets* publiés par l'utilisateur n pendant la période étudiée.

J'ai choisi de travailler avec la citation comme interaction pour plusieurs raisons. D'abord, Boyd, Golder et Lotan [25] montrent dans leur étude datant de 2010 et depuis largement relayée et admise par la communauté scientifique que le *retweet*, par sa nature de fonctionnalité de partage, est une interaction impliquant la plupart du temps une approbation ou un soutien, ce qui est donc incompatible avec la mesure d'antagonisme. Conover et al. [37] étudient les graphes des mentions issus de Twitter et concluent qu'ils ne sont en général pas polarisés. Enfin, Guerra annonce dans un article de 2017 [59] que la meilleure interaction de Twitter à utiliser avec sa méthode est la citation, qui est beaucoup utilisée à des fins d'humour et de sarcasme pour critiquer ou détourner l'idée exprimée de base, ce qui favorise l'antagonisme.

Le graphe G_Q a été nettoyé en suivant les indications formulées par Garimella et al. [50] et souvent reprises dans la littérature, qui préconisent de retirer les interactions occasionnelles en supprimant les arêtes avec un poids égal à 1 pour réduire le bruit présent dans le réseau. Le graphe final G_Q contient 24 591 sommets et 55 703 arêtes. Les degrés entrants et sortants moyens sont égaux à 2,265.

2.5.2 Analyse des communautés

Pour détecter les communautés d'utilisateurs dans G_Q, j'ai choisi d'utiliser l'algorithme de Louvain [18]. Ce choix a été motivé par mes expériences passées sur des études de cas similaires menées dans le cadre de mon projet tuteuré et au cours desquelles cet algorithme en particulier avait permis d'obtenir les structures communautaires les plus intéressantes à commenter car moins diluées à l'intérieur d'un grand nombre de très petites communautés. J'ai ainsi obtenu 8 communautés avec une taille suffisante pour être considérées comme significatives, c'est-à-dire avec une taille supérieure à la limite de résolution du graphe [53], qui est égale à 333. La modularité [113] du graphe est égale à 0.59, G_Q a donc une structure communautaire forte composée de groupes de sommets avec une grande cohésion.

Pour faciliter l'interprétation des résultats, j'ai manuellement étiqueté ces 8 communautés en étudiant les *hashtags* les plus utilisés (*top-hashtags*) dans les publications de leurs membres. Sur Twitter, un *hashtag* est une annotation manuelle qu'un utilisateur peut placer à l'intérieur de sa publication en préfixant le mot-clé qu'il souhaite par le symbole #. Pour récupérer ces *top-hashtags*, j'ai construit pour chaque communauté un graphe biparti *user-hashtag* contenant d'une part les utilisateurs membres de la communauté et d'autre part les *hashtags* qu'ils ont déjà utilisés. J'ai ensuite calculé les 30 sommets les plus centraux selon leur degré entrant de la partie *hashtags* de chaque graphe biparti. Pour finir, j'ai retiré des listes obtenues les termes neutres qui même avec le contexte fourni par les autres ne permettent pas de donner une indication sur l'opinion partagée par le groupe, par exemple les noms de figures publiques (Macron, Castex, etc.), de vaccins (Pfizer, Moderna, etc.) et de laboratoires (BioNTech, etc.). Les 8 communautés, avec leur taille et leurs *top-hashtags* respectifs, sont présentées dans la table 2.1.

On remarque la présence de deux communautés principales qui contiennent à elles deux plus de 60% des utilisateurs. Leurs thématiques principales s'articulent autour de la problématique de la vaccination, une dans un sens positif et l'autre dans un sens négatif. On peut donc déjà conclure sur la présence de deux pôles au sein de la structure communautaire, qui d'un point de vue sémantique semblent opposés. Dans la suite de l'étude de cas, l'analyse se concentre uniquement sur ces deux communautés.

Des discours inter-communautaires sont identifiables dans les *top-hashtags*. Le premier et plus fréquent est un discours anti-gouvernemental, représenté par des *hashtags* tels que "Dicta-

ID	Taille	Catégorie	Top-hashtags nettoyés (sans #)
20792	10 604	Pro-vaccins	AFP, Mutation, Confinement3, CouvreFeu, Ecoles, EHPAD, PasseportVert, DictatureSanitaire, Israël, JeMeFaisVacciner, Pasteur
12884	4 722	Anti-vaccins	Ivermectine, DictatureSanitaire, JeNeMeConfineraiPas, Raoult, Hydroxychloroquine, EtLesSoins, Plandémie, VéranDémission, LesPierresCrieront, GreatReset, Ethique, BeBraveWHO, JeNeMeVaccineraiPas
17387	2 736	Réactions médias	AFP, Confinement3, DictatureSanitaire, CouvreFeu, SudRadio, Le79Inter, Ivermectine, Santé, EDPHAD, LCI, Cnews, ZeroCovid, La26, PasseportVaccinal, Variant
11672	2 638	Anti-Blanquer	BlanquerMent, BlanquerDemission, Blanquer, PasDeVague, CovidLong, Ecoles, GarderieNationale, ProtocoleFantome, ProtocoleBidon, ParentsEnColere
16302	1 649	Politique	LR, LCI, UE, RN, DictatureSanitaire, OlivierVeran, LFI, PS, EmmanuelMacron, Medias, LREM, Melenchon
16079	976	Québecois	PolQC, Covid19qc, PolCan, CAQ, Québec, PLQ, QCPolic, EduQC, Montreal
15179	882	Anti-gouvernement	LREM, StopDictatureSanitaire, Macronie, JeSuisLibre, MacronDestitution, RéveillezVous, TousContreMacron, BlanquerDemission, Gouvernement
14931	384	Sans étiquette	Raoult, LREM, Ivermectine, DictatureSanitaire, PasseportVaccinal, Macronie, JeNeMeConfineraiPas, Trump, Cnews, VeranDemission

TABLE 2.1 – Communautés significatives de G_Q découvertes par l'algorithme de Louvain.

tureSanitaire", "JeNeMeConfineraiPas", "VéranDémission", "BlanquerDémission" ou encore "MacronDestitution". Le deuxième est un discours complotiste, représenté par des *hashtags* comme "Plandémie" [4], "GreatReset" [5] ou "BlanquerMent". Ces deux thématiques semblent donc suivre celle de la vaccination contre la Covid-19 dans la plupart des conversations, et ce quelle que

4. Théorie du complot qui stipule que "la pandémie de Covid-19 serait un plan visant à imposer un "nouvel ordre mondial" pour asservir les populations et réduire le nombre d'habitants sur le globe." `https://factuel.afp.com/plandemie-un-texte-viral-contenant-de-fausses-informations`

5. Théorie du complot détournée d'un livre du même nom rédigé par deux membres du Forum économique mondial de Davos qui stipule que la pandémie "consisterait à vouloir détruire l'économie et les petits commerces, supprimer la monnaie, mettre à bas les démocraties, imposer un suivi personnalisé de bonne citoyenneté à la chinoise ou encore mettre en place un régime tantôt communiste, fasciste, nazi ou les trois à la fois." `https://www.lemonde.fr/les-decodeurs/article/2021/02/10/qu-est-ce-que-the-great-reset-un-livre-devenu-theorie-du-complot_6069491_4355770.html`

soit la thématique principale de la communauté. Cette observation peut être interprétée comme un signe de la grande dimension politique qu'a pris la résolution de la crise, au-delà de l'aspect sanitaire, qui s'accompagne d'une forme de méfiance vis-à-vis des hautes autorités.

2.5.3 Analyse de la polarisation

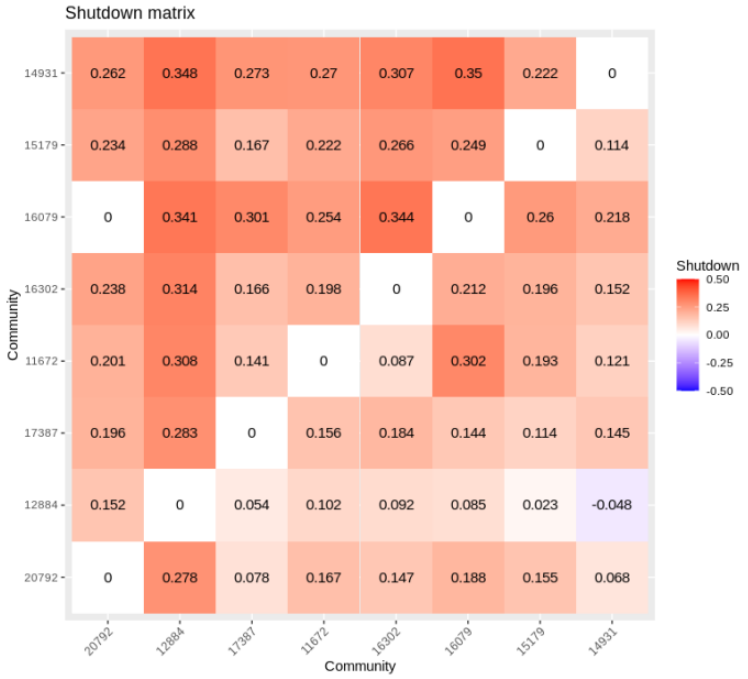
Une fois les communautés identifiées, j'ai calculé les mesures de polarisation de G_Q. La figure 2.5 montre les résultats obtenus. Pour rappel, la communauté pro-vaccins possède l'identifiant 20792 et la communauté anti-vaccins l'identifiant 12884. Sur la matrice d'antagonisme 2.5a, ces communautés sont représentées sur les premières colonnes et sur les dernières lignes.

Par rapport à la communauté anti-vaccins, la colonne 12884 résume le comportement de ses frontières avec les autres communautés. On remarque d'abord que cette colonne concentre quasiment la totalité des plus hauts scores d'antagonisme de la matrice. On a alors ici affaire à une communauté très renfermée, où les membres frontières sont très investis à l'intérieur du groupe, et qui est donc assez susceptible d'être antagoniste avec la totalité des autres communautés de l'environnement social. Sur la ligne 12884, on peut observer le degré d'antagonisme des autres communautés avec les anti-vaccins. Les scores sont cette fois-ci très bas. Les autres communautés interagissent donc pas mal avec la communauté anti-vaccins, quasiment autant qu'avec l'intérieur de leur propre communauté. On remarque alors une forme d'investissement de la part des autres à vouloir répondre aux membres de la communauté anti-vaccins, dans le meilleur des cas à des fins de pédagogie et dans le pire à des fins d'acharnement. Dans tous les cas, les scores ne traduisent pas vraiment une forte probabilité que les frontières soient antagonistes. On peut toutefois noter que le score le plus élevé correspond à celui de la communauté pro-vaccins, qui est donc la plus susceptible d'être antagoniste avec les anti-vaccins.

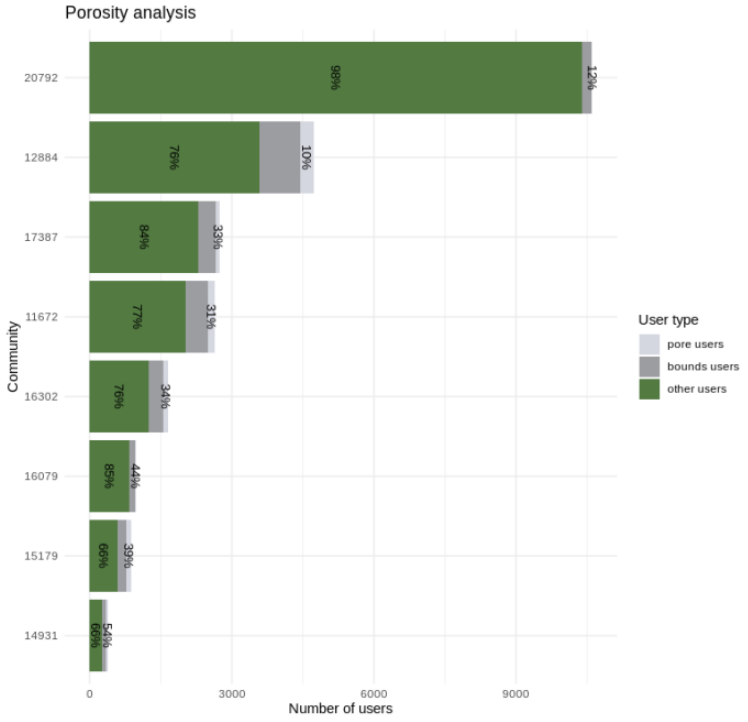
Concernant les pro-vaccins (identifiant 20792), en plus de la remarque faite précédemment, on observe des valeurs moins élevées sur la colonne par rapport à celles vues avec les anti-vaccins, et inversement sur la ligne. Cette communauté serait donc moins susceptible d'attaquer les autres mais serait prise à partie par des interventions plus brèves et donc possiblement plus antagonistes. La communauté qui est la plus susceptible d'être antagoniste avec les pro-vaccins est celle des anti-vaccins. Avec une analyse purement structurelle, on remarque alors une nouvelle fois une forme de rivalité mutuelle.

Pour finir, les scores de porosité des frontières de la figure 2.5b permettent de commenter la cohésion des deux pôles. On remarque que la communauté pro-vaccins est formée de très petites frontières, dont les membres ne représentent que 2% de la totalité des individus de la communauté, très renfermées sur elles-mêmes. En effet, malgré sa grande taille en comparaison des autres communautés de l'environnement, seuls 12% des membres frontières citent plus ou autant de contenu en provenance d'utilisateurs extérieurs à la communauté qu'en provenance d'utilisateurs internes.

Au contraire, la communauté anti-vaccins est formée de plus grandes frontières, dont les membres représentent environ 24% de la totalité des utilisateurs de la communauté. Cependant, celles-ci restent aussi très renfermées, avec seulement 10% de membres frontières qui possèdent un score d'antagonisme négatif ou nul. Par conséquent, ces interventions, bien qu'orchestrées par un plus grand nombre d'individus uniques, restent brèves et loin d'être majoritaires par rapport à celles concernant des *tweets* publiés par des membres de la communauté anti-vaccins. Dans les deux cas, ces comportements sont signes de chambres d'échos, où les individus s'enferment intellectuellement et d'un point de vue informationnel en n'étant confrontés plus qu'à du contenu



(a) matrice d'antagonisme



(b) porosité des frontières

FIGURE 2.5 – Indicateurs de polarisation calculés par application de la méthode sur G_Q.

qui les conforte dans leur vision du monde et dans leurs convictions, au point de devenir incapable de porter un regard critique et constructif sur le monde qui les entoure.

2.5.4 Conclusion de l'étude de cas

Dans cette étude de cas, j'ai décidé d'analyser un graphe de citations (fonctionnalité de partage avec ajout d'un commentaire) extrait d'un jeu de données collecté sur Twitter et traitant des thématiques de la vaccination et de la Covid-19. Grâce à l'algorithme de détection de communautés Louvain, j'ai découvert 8 communautés significatives construites autour de thématiques et opinions différentes. Parmi elles, on retrouve notamment deux pôles principaux aux tendances pro et anti-vaccins. La valeur de modularité assez élevée du graphe indique que ces 8 communautés présentent une forte cohésion. Après application de la méthode présentée dans ce document, les valeurs de porosité des frontières confirment cette observation et illustrent un comportement particulier de renfermement chez les deux plus grosses communautés, qui peut être vu comme un signe de formation de chambres d'échos. Les valeurs d'antagonisme indiquent que les deux communautés pro et anti-vaccins sont très susceptibles d'être mutuellement antagonistes.

En 1986, Isenberg a défini la polarisation sociale comme le processus menant un groupe d'individus à se scinder en deux sous-groupes opposés possédant des opinions contrastées et conflictuelles sur un sujet donné, avec peu d'individus qui restent neutres ou dans une position intermédiaire [73]. Grâce aux informations collectées dans cette étude de cas, on sait que l'environnement social étudié se scinde en 8 sous-groupes, mais que parmi eux 2 se démarquent par leur taille (plus de 60% du total des individus). Ces deux sous-groupes s'opposent autour de la question de la vaccination en défendant chacun un avis tranché sur la question (pour/contre), le tout de manière assez conflictuelle. On peut donc conclure que le graphe des citations révèle une polarisation des individus sur Twitter par rapport à la problématique de la vaccination contre la Covid-19.

2.5.5 Bonus : analyse du discours des frontières

Un autre cas d'utilisation possible de la méthode, qui utilise le découpage en zones internes et frontières pour étudier et commenter les stratégies de communication des communautés, est présenté dans cette sous-section. Ces travaux se situent un peu à part, à la fois de la présentation de la méthode et de l'étude de cas, car elle présente un détournement de l'approche pour atteindre un autre objectif que la détection de polarisation et car elle n'a pour le moment pas été expérimentée sur d'autres jeux de données. Toutefois, puisque les résultats sont intéressants, je lui dédie une partie en bonus.

J'ai commencé par réaliser une expérience qui consistait à récupérer les 20 *hashtags* les plus utilisés par les frontières des communautés pro et anti-vaccins lors de leurs interactions avec les autres communautés du graphe. Les listes obtenues sont présentées dans les tables 2.2 et 2.3. Les *hashtags* qui font partie des *top-hashtags* de la communauté "Frontière avec ..." sont représentés en orange. Ceux qui font partie des *top-hashtags* de la communauté pro ou anti-vaccins, selon le tableau, en **gras**. Enfin, ceux qui apparaissent dans les *top-hashtags* de toutes ou presque toutes les frontières sont soulignés.

Frontière avec ...	**Top-hashtags (sans #)**

Anti-vaccins	**Confinement3**, **CouvreFeu**, DictatureSanitaire, GreatReset, Raoult, **Astrazenaca**, BigPharma, Ivermectine, JeNeMeConfineraiPas, CouvreFeu18h, **AFP**, HydroxyChloroquine, PasseportVaccinal, StopDictatureSanitaire, GrandReset, **Santé**, HdPros, Grippe, Thread, Remdesivir
Réactions médias	**AFP**, **Confinement3**, MediaCitoyens, GGRMC, **Blanquer**, Nice06, **CouvreFeu18h**, **Astrazenaca**, BlanquerDemission, **CouvreFeu**, **Allemagne**, **Castex18h**, Sanofi, BlanquerMent, HdPros, Reconfinement, DictatureSanitaire, Soignants, **Ehpad**, **Santé**
Anti-Blanquer	**Blanquer**, GreatReset, **AFP**, **Confinement3**, BlanquerDemission, Israel, **Astrazenaca**, GrandReset, **CouvreFeu**, Pikouze, NWO, **Santé**, BlanquerMent, Reconfinement, Grippe, BigPharma, Holdup, **Castex18h**, GGRMC, Covac40
Politique	**Blanquer**, **Astrazenaca**, **Confinement3**, **AFP**, BlanquerDemission, **Castex18h**, **CouvreFeu**, BlanquerMent, GGRMC, TPMP, BFMTV, CNews, Remdesivir, Raoult, FranceInter, **CouvreFeu18h**, Sputnikv, PasseportVaccinal, Reconfinement, **Allemagne**
Québecois	Pas de frontière.
Anti-gouvernement	**Confinement3**, Nice06, Raoult, **CouvreFeu**, BigPharma, Ivermectine, **CouvreFeu18h**, **AFP**, DictatureSanitaire, **Astrazenaca**, Spoutnikv, Plandemie, Thread, Sputnikv, AlpesMaritimes, **Italie**, JeNeMeConfineraiPas, Reconfinement, **Allemagne**, **Santé**
Sans étiquette	GreatReset, GrandReset, Pikouze, NWO, HoldUp, Grippe, Covac40, BillGates, Orwell, DroitsDelHomme, Citoyens, Peuple, INSEE, Lobbys, Sida, FEMA, TV, AI, Attali, BigPharma

TABLE 2.2 – Tableau de la communauté pro-vaccins.

Frontière avec ...	**Top-hashtags (sans #)**
Pro-vaccins	**EtLeSoin**, **DictatureSanitaire**, **JeNeMeConfineraPas**, **Ivermectine**, **BigPharma**, **StopDictatureSanitaire**, **StopCouvreFeu**, **GreatReset**, **Raoult**, **VeranDemission**, Macronie, **Confinement3**, **Hydroxychloroquine**, **Plandemie**, ReveillezVous, **CouvreFeu**, DeepStateCorruption, Assassins, Petits-Commerces, Desobeissez
Réactions médias	**EtLeSoin**, **Ivermectine**, **DictatureSanitaire**, **BigPharma**, **StopDictatureSanitaire**, **JeNeMeConfineraiPas**, **Raoult**, **Hydroxychloroquine**, **StopCouvreFeu**, **CouvreFeu**, **VeranDemission**, **Confinement3**, **PasseportVaccinal**, **Plandemie**, Thread, **Ethique**, **Masques**, **LesPierresCrieront**, Astrazenaca, JeNeMeVaccineraiPas
Anti-Blanquer	**EtLeSoin**, **Ivermectine**, **Raoult**, **BigPharma**, **DictatureSanitaire**, **JeNeMeConfineraiPas**, **Hydroxychloroquine**, **StopDictatureSanitaire**, **CouvreFeu**, Thread, **StopCouvreFeu**, **VeranDemission**, **Confinement3**, **PasseportVaccinal**, **Plandemie**, **Remdesivir**, HCQ, **GreatReset**, Santé, **LesPierresCrieront**

Politique	**EtLeSoin**, **Ivermectine**, **BigPharma**, **DictatureSanitaire**, **JeNeMeConfineraiPas**, **StopDictatureSanitaire**, **Raoult**, **Hydroxychloroquine**, **StopCouvreFeu**, **VeranDemission**, **CouvreFeu**, **Confinement3**, **Plandemie**, **PasseportVaccinal**, **GreatReset**, HCQ, **LesPierresCrieront**, **Remdesivir**, **Masques**, **Ethique**
Québecois	**DictatureSanitaire**, **JeNeMeConfineraiPas**, **Ivermectine**, **GreatReset**, **BigPharma**, ReveillezVous, **StopCouvreFeu**, **StopDictatureSanitaire**, DeepStateCorruption, Assassins, PetitsCommerces, Desobeissez, PolQC, **Plandemie**, **Hydroxychloroquine**, **Raoult**, Fossoyeurs, **Confinement3**, **VeranDemission**, **CouvreFeu**
Anti-gouvernement	**EtLeSoin**, **Ivermectine**, **DictatureSanitaire**, **BigPharma**, **JeNeMeConfineraiPas**, **Raoult**, **StopDictatureSanitaire**, **StopCouvreFeu**, **Hydroxychloroquine**, **VeranDemission**, **CouvreFeu**, **Confinement3**, **GreatReset**, **Plandemie**, **PasseportVaccinal**, HCQ, **Masques**, **Ethique**, **Remdesivir**, **LesPierresCrieront**
Sans étiquette	**DictatureSanitaire**, **JeNeMeConfineraiPas**, **Ivermectine**, **BigPharma**, **GreatReset**, ReveillezVous, **StopCouvreFeu**, **StopDictatureSanitaire**, **Raoult**, Assassins, DeepStateCorruption, PetitsCommerces, Desobeissez, **VeranDemission**, **Hydroxychloroquine**, Fossoyeurs, **CouvreFeu**, **PasseportVaccinal**, **Confinement3**, Escroquerie

TABLE 2.3 – Tableau de la communauté anti-vaccins.

Ensuite, j'ai identifié dans chaque liste les *top-hashtags* communs entre la frontière et la deuxième communauté de la paire (celle qui est mentionnée dans la colonne de gauche "Frontière avec ..."), représentés en orange dans les tableaux. L'idée était d'identifier les thématiques que les membres des frontières pro ou anti-vaccins se sont appropriés lors de leurs interactions avec les autres communautés. En moyenne, 6.7 *top-hashtags* sur 20 (33.5%) d'une zone frontière de la communauté pro-vaccins font aussi partie de la liste des *top-hashtags* de la communauté avec laquelle elle interagit, contrairement à seulement 4.4 *top-hashtags* sur 20 (22%) pour une zone frontière de la communauté anti-vaccins. Quand on s'intéresse de près aux *top-hashtags* communs, on remarque également une nouvelle différence. En effet, alors que quelques *hashtags* spécifiques aux thématiques de l'autre communauté de la paire sont utilisés dans le tableau 2.2, comme "GreatReset" et "JeNeMeConfineraiPas" dans la frontière avec les anti-vaccins ou "BlanquerMent" et "BlanquerDemission" dans celle avec les anti-Blanquer, les termes oranges dans le tableau 2.3 restent globalement centrés sur les thématiques propres aux anti-vaccins, comme par exemple "Ivermectine", "JeNeMeConfineraiPas" ou "Raoult".

J'ai alors décidé d'identifier les *top-hashtags* communs entre les zones frontières et leur propre communauté, marqués en gras dans les tableaux. On remarque ainsi que là où en moyenne 6.3 *top-hashtags* sur 20 (31.5%) d'une zone frontière de la communauté pro-vaccins appartiennent aussi aux *top-hashtags* de leur communauté, 16 *top-hashtags* sur 20 (80%) d'une zone frontière de la communauté anti-vaccins appartiennent aussi aux *top-hashtags* de la leur. De plus, il s'avère que sur les 4.4 *top-hashtags* communs entre les frontières anti-vaccins et les autres communautés des paires, 4.3 font aussi partie de la liste des *top-hashtags* de la communauté anti-vaccins (termes en orange gras). Cela signifie qu'en moyenne, 97.7% des *top-hashtags* communs entre une frontière anti-vaccins et la communauté avec laquelle elle interagit sont en réalité des *top-hashtags* de la communauté anti-vaccins. Leur utilisation n'induit donc pas d'adaptation de la

part des utilisateurs des frontières anti-vaccins aux thématiques propres à l'autre communauté de la paire, puisque celles-ci sont très majoritairement des discours inter-communautaires. Du côté des frontières pro-vaccins, ce chiffre s'élève à seulement 3.2 *top-hashtags* sur 20, soit 47.7%, ce qui traduit une plus forte adaptation.

Enfin, ma dernière expérience consistait à construire un histogramme des *top-hashtags* des frontières de chacune des deux communautés. J'ai ainsi découvert que dans le cas de la communauté pro-vaccins, aucun *hashtag* n'est présent dans les *top-hashtags* de toutes les frontières. Seuls 4 termes, "AFP", "Astrazenaca" (avec la faute), "CouvreFeu" et "Confinement3", soulignés dans le tableau 2.2, font partie des *top-hashtags* de 5 frontières sur les 6. En tout, 8 termes appartiennent aux *top-hashtags* de plus de la moitié des frontières de la communauté pro-vaccins (voir figure 2.6). Alors que dans le cas de la communauté anti-vaccins, 10 termes sont utilisés dans la totalité des frontières : "CouvreFeu", "Confinement3", "VeranDemission", "StopDictatureSanitaire", "Ivermectine", "DictatureSanitaire", "BigPharma", "StopCouvreFeu", "Raoult" et "Hydroxychloroquine". On retrouve 16 termes dans les *top-hashtags* de plus de la moitié des frontières de la communauté anti-vaccins (voir figure 2.7). Un résumé des valeurs numériques identifiées dans cette partie est disponible dans la table 2.4.

FIGURE 2.6 – *Top-hashtags* uniques présents dans plus de la moitié des frontières de la communauté pro-vaccins.

De plus, alors que les *top-hashtags* fréquents des frontières pro-vaccins restent globalement très neutres, ceux des frontières anti-vaccins sont très spécifiques aux thématiques principales de cette communauté, c'est-à-dire les traitements alternatifs ("Ivermectine", "Hydroxychloroquine", "Raoult", "EtLeSoin[6]"), les critiques anti-gouvernementales ("VeranDemission", "StopDictatureSanitaire", "StopCouvreFeu", "JeNeMeConfineraiPas") et les théories du complot ("Plandemie", "GreatReset"). Finalement, on peut noter que l'Agence France Presse (AFP) se trouve à chaque fois dans les premiers *top-hashtags* des frontières pro-vaccins (dans 5 listes sur 6), alors qu'elle est pourtant complètement absente de tous les *top-hashtags* de la communauté anti-vaccins. On peut ou bien interpréter ici un manque de confiance envers cette entité, ou alors le signe d'une propension moins importante à partager des informations en provenance de sources officielles.

Pour faire une synthèse des résultats précédents et fournir quelques interprétations supplémentaires, on observe une différence flagrante entre les stratégies de communication des com-

6. Dénonce l'absence d'intérêt pour la recherche de traitements au profit des solutions préventives.

FIGURE 2.7 – *Top-hashtags* uniques présents dans plus de la moitié des frontières de la communauté anti-vaccins.

munautés pro et anti-vaccins avec les autres. Les frontières de la communauté pro-vaccins ont tendance à adapter leurs discours aux communautés avec lesquelles elles interagissent, en utilisant approximativement la même proportion de *hashtags* communs avec leur communauté qu'avec l'autre. Par conséquent, peu de termes sont majoritairement utilisés dans tous les cas de figure, et les plus fréquents sont ou bien très généraux et neutres, relevant ainsi du discours inter-communautaire, ou bien relatifs au relais de sources officielles comme l'AFP. Au contraire, les frontières de la communauté anti-vaccins cristallisent leurs communications autour d'un ensemble fixe et bien connu de *hashtags* et n'emploient pas ou peu ceux des autres communautés.

Tous ces éléments indiquent que la communauté pro-vaccins est moins organisée et plus offensive, dans le sens où elle hésite moins à aller confronter l'avis des autres "sur leur terrain". Au contraire, la communauté anti-vaccins est bien plus organisée et défensive. Elle se renferme sur elle-même en permettant aux membres de la communauté de se retrouver entre eux [7] et passe la majeure partie de son temps d'exposition à l'extérieur à défendre l'opinion du groupe. Si on compare cette conclusion avec les évaluations de l'antagonisme, les valeurs moins élevées de la colonne des pro-vaccins vont dans le sens d'interactions plus fréquentes avec l'extérieur, ce qui est cohérent avec la théorie de la communauté offensive. De même, les valeurs d'antagonisme très élevées de la colonne des anti-vaccins sont un signe de frontières beaucoup plus investies vers l'intérieur de la communauté, ce qui, une fois de plus, va dans le sens de la communauté défensive et organisée.

7. Quand un utilisateur de Twitter publie un message contenant un *hashtag* particulier, n'importe qui peut facilement retrouver son *tweet* sans connaître personnellement l'auteur grâce à une simple recherche du terme sur le RSN.

Mesure	**Pro-vaccins**	**Anti-vaccins**
Nombre moyen de *top-hashtags* communs entre une frontière et l'autre communauté de la paire	6.7/20 (33.5%)	4.4/20 (22%)
Proportion des *top-hashtags* communs entre une frontière et l'autre communauté qui **font aussi partie** des *top-hashtags* de sa communauté	3.2/6.7 (47.7%)	4.3/4.4 (97.7%)
Nombre moyen de *top-hashtags* communs entre une frontière et **sa propre communauté**	6.3/20 (31.5%)	16/20 (80%)
Nombre de *top-hashtags* uniques présents dans toutes les frontières de la communauté	0	10
Nombre de *top-hashtags* uniques présents dans plus de la moitié des frontières de la communauté	8	16

TABLE 2.4 – Tableau récapitulatif des valeurs numériques.

2.6 Conclusion et perspectives

Pour conclure ce chapitre, j'ai travaillé au cours de mon Master parcours recherche sur une nouvelle contribution à la détection de polarisation au sein des graphes sociaux. L'approche que je propose est uniquement basée sur une analyse de la structure du graphe, plus précisément des frontières de ses communautés, dans le but de développer une méthode complètement automatique, générique et non-supervisée. Ma contribution peut être considérée comme une extension des travaux de GUERRA et al. [58], plus adaptée aux données réelles grâce à la prise en considération de la direction des liens, de leur pondération et des structures communautaires recouvrantes. L'identification et l'analyse des frontières des communautés permet le calcul d'indicateurs utiles à la détection de polarisation, tels que la matrice d'antagonisme ou la porosité des frontières, mais aussi l'analyse de discours. En plus de la méthode elle-même, est également proposée l'implémentation en R d'un algorithme disponible en libre accès. La dernière partie de ce chapitre est constituée d'une étude de cas réelle menée dans le cadre du projet interdisciplinaire Cocktail.

Comme mentionné à plusieurs reprises tout au long de ce mémoire, mon travail sur la polarisation a concerné à la fois mon projet tuteuré et la première moitié de mon stage. Concrètement, voici les éléments qui étaient déjà présents à la fin du projet tuteuré :

- la méthode en elle-même avec une première version de l'implémentation en R, sans la gestion des communautés recouvrantes ;
- une première version moins aboutie de l'état de l'art ;
- une première étude de cas, complètement différente de celle présentée dans ce document ;
- un article de recherche de 16 pages en français accepté pour publication par le comité de lecture de la conférence INFORSID 2021.

Dans le cadre de la première moitié du stage, les éléments suivants ont abouti :

- modification de la méthode et de son implémentation pour gérer les communautés recouvrantes ;
- visualisation des résultats de la méthode (*heat map* pour la matrice d'antagonisme, *bar chart* pour la porosité des frontières) ;

— rédaction d'un deuxième article court de 4 pages en anglais accepté pour publication par le comité de lecture de la conférence FRCCS 2021 ;

— étude de cas présentée dans ce document ;

— préparation et présentation d'exposés aux conférences INFORSID et FRCCS ainsi que lors d'un séminaire du LIB ;

— rédaction d'un troisième article de 10 pages en anglais à destination d'une conférence internationale de rang B, terminé à 90% mais dont la rédaction a dû être mise en pause à cause de la période estivale et pour permettre la rédaction de ce mémoire.

Certains autres éléments ont également été travaillés lors du stage mais n'ont pas abouti sur des résultats assez intéressants et/ou complets pour être présentés dans ce document. Ils font donc à l'heure actuelle partie des perspectives de ce travail.

Premièrement, j'ai eu l'occasion de travailler en fond sur des optimisations de l'algorithme et de son implémentation en R. Après quelques réflexions, trois approches sont pour moi possibles aujourd'hui pour réduire sa complexité et améliorer ses performances : 1) supprimer la deuxième étape de la construction de la matrice structurelle, ce qui reviendrait à juste identifier les membres des zones internes puis de se servir de cette liste pour identifier les membres des zones frontières et calculer leur score d'antagonisme en même temps ; 2) explorer la piste d'un calcul matriciel comme peuvent le proposer certains algorithmes tels que PageRank ; 3) paralléliser l'algorithme. J'ai eu l'occasion de travailler sur ces trois pistes, comme les branches de mon GitHub peuvent en attester, mais le temps m'a manqué pour aboutir sur un résultat final.

Deuxièmement, il faudrait approfondir les possibles corrélations entre les indicateurs développés dans ma contribution et ceux traditionnellement utilisés en SNA, comme la modularité ou les différentes centralités. Au début du stage, plusieurs hypothèses ont été avancées et seule la première a pu été traitée : Les zones frontières d'une communauté contiennent majoritairement des hubs, les zones internes des autorités. Une analyse des quelques expériences menées sur la question est disponible dans l'annexe A.1. En résumé, les résultats des expériences ne permettent pas de confirmer ou d'infirmer l'hypothèse. Les autres hypothèses que je n'ai pas eu le temps d'approfondir sont les suivantes :

— Les communautés qui possèdent des individus centraux dans leurs frontières possèdent de plus grosses zones frontières.

— Plus la centralité moyenne d'une zone frontière est faible, plus l'antagonisme sera fort.

— Plus la modularité de la communauté est élevée, plus ses frontières seront vecteurs d'antagonisme.

— Plus la modularité d'un graphe est grande, plus la porosité des frontières sera faible.

Enfin et pour finir, un aspect qui a finalement été exploré assez rapidement et qui mériterait probablement plus d'attention est l'étude des discours des zones internes et frontières dans le but de mieux comprendre le comportement des communautés. Les premiers résultats obtenus sont très encourageants et permettent de souligner une réelle divergence entre les stratégies de communication des pro et des anti-vaccins. Certains indicateurs numériques pourraient même être envisagés comme le nombre de *hashtags* uniques utilisés par toutes les frontières, ou encore la proportion de *top-hashtags* communs avec chacune des deux communautés de la paire. Ils pourraient alors donner des indications concernant le degré d'organisation des frontières et concernant leur centralisation autour de termes particuliers.

Chapitre 3

Enrichissement des données issues des réseaux sociaux numériques par annotation sémantique transmédia

3.1 Introduction

Ces dernières années, l'analyse des réseaux sociaux numériques (RSN) s'est hissée parmi les problématiques principales des entreprises et acteurs de la recherche. Son enjeu majeur est de réussir à comprendre le comportement, les sentiments et les opinions d'un ensemble d'individus à partir des données qu'ils produisent sur des plateformes comme Facebook, Twitter, etc.

Une contrainte régulièrement rencontrée lors de l'analyse des messages publiés sur les RSN est leur taille. Sur Twitter par exemple, chaque publication ne peut dépasser 280 caractères, ce qui force l'expression d'avis et de commentaires moins nuancés, moins justifiés et moins documentés. Pour pallier le manque d'espace à leur disposition pour correctement s'exprimer, les utilisateurs qui en ressentent le besoin ont pour habitude d'utiliser deux stratégies principales :

— utiliser les fonctionnalités multimédias proposées par Twitter comme l'import d'images ;

— partager des liens vers des médias hébergés à l'extérieur du RSN, comme par exemple des pages HTML, des documents PDF, des vidéos, etc. ;

Sur le jeu de données construit pour l'étude de cas de la section 2.5 du chapitre précédent, 10.9% des *tweets*[1] utilisent la première stratégie, 31.5% le deuxième et 38% au moins l'un des deux. Avec des méthodes adéquates dites d'annotation sémantique, il serait donc possible d'enrichir le sens de plus d'un *tweet* sur trois en prenant en considération les médias référencés et en résumant leur contenu par un ensemble de mots-clés.

Pour les médias importés avec les fonctionnalités de Twitter, la figure 3.1 montre leurs types et extensions. Puisque le RSN ne fournit que les miniatures (première image) des médias, qu'ils soient vidéos, animés ou figés, tous les fichiers à traiter sont des images aux formats JPG ou PNG. Pour les médias externes, la catégorisation est plus complexe à cause de l'absence d'un modèle unique pour les URL, contrairement aux médias importés, qui rend leur traitement difficile à l'aide d'expressions régulières. Cependant, la syntaxe commune inhérente à toutes les

1. Par rapport à l'ensemble des *tweets* qui ne sont pas des *retweets* (originaux, réponses, citations).

URL permet de détecter les noms de domaines qui hébergent les médias référencés (figure 3.2). En étudiant de plus près les résultats, on retrouve principalement :

— la plateforme d'hébergement vidéo Youtube (à deux reprises) ;

— des services qui permettent de raccourcir les URL comme *bit.ly* ou de programmer des *tweets* comme *ift.tt* et *divr.it*, pour lesquels il est impossible de connaître le type du média référencé sans le télécharger ;

— des sites d'information, qui hébergent donc probablement des pages HTML.

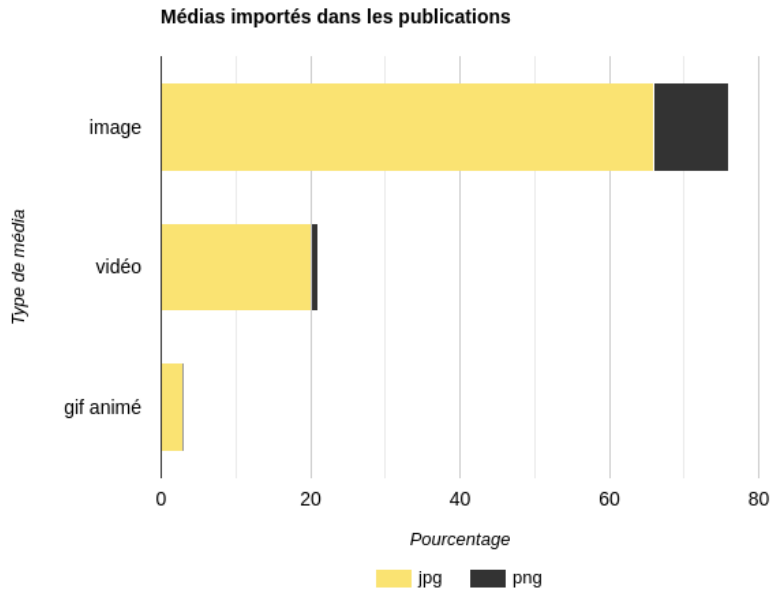


FIGURE 3.1 – Types et extensions de médias importés dans les *tweets*.

Après analyse du jeu de données, on peut donc extraire trois grandes catégories de médias utilisés pour enrichir les *tweets* : les images, les documents textuels non ou semi-structurés (notamment pages HTML) et les vidéos. Dans le cadre de mon stage, j'ai décidé de d'abord concentrer mes efforts sur l'annotation sémantique des documents textuels, plus nombreux et plus riches sémantiquement.

L'affectation et l'extraction de mots-clés sont des problématiques très explorées en sciences des données, et ce depuis très longtemps, grâce à leur potentiel pour réduire la complexité de traitement et d'analyse des données non-structurées en les représentant par un ensemble fini de termes, identifiés selon leur capacité à résumer ou décrire le sens de la donnée de départ. Différemment, les deux méthodes permettent d'annoter sémantiquement des documents.

Dans le cas des données non-structurées textuelles, les méthodes d'extraction de mots-clés (en anglais *Keyword Extraction* ou KE), où tous les termes-clés sont explicitement mentionnés dans les données, sont plutôt privilégiées. Dans le cas des données non-structurées de type images, ce sont les méthodes d'affectation de mots-clés (en anglais *Keyword Assignment* ou KA), où les termes-clés sont tous choisis à partir d'un vocabulaire contrôlé ou d'une taxonomie prédéfinie, qui sont privilégiées. Les méthodes de KA sont également parfois utilisées sur les données textuelles,

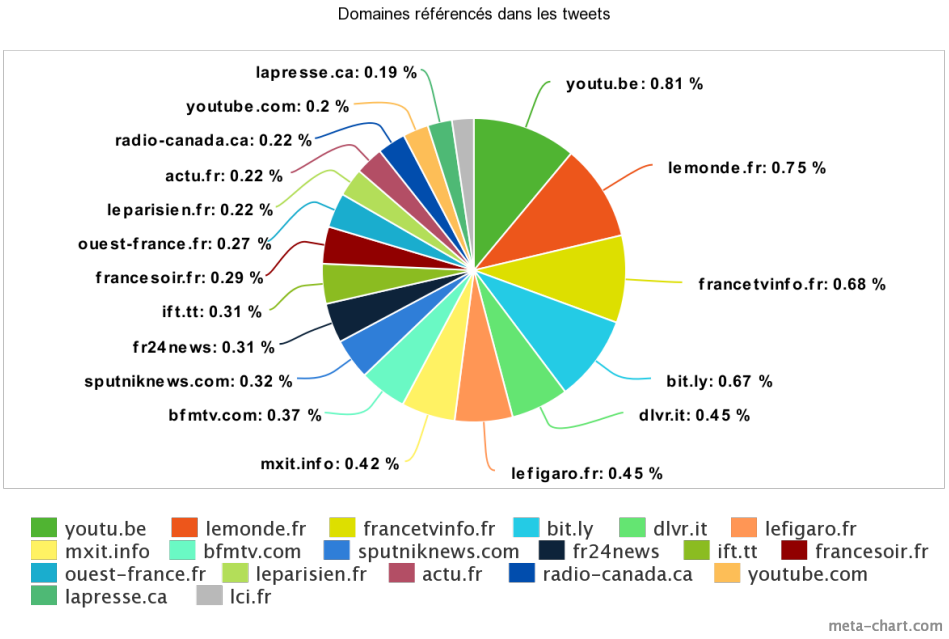


FIGURE 3.2 – 20 domaines les plus référencés dans les URL incluses dans les *tweets*. Les valeurs correspondent aux pourcentages de représentation parmi l'ensemble des domaines référencés dans le jeu de données.

mais la limite de description imposée par la définition du vocabulaire constitue une contrainte généralement jugée non-nécessaire quand les textes étudiés sont sémantiquement assez riches.

Dans la suite du chapitre, la section 3.2 dresse un état de l'art sur l'extraction automatique de mots-clés. La section 3.3 présente un comparatif de différentes approches, conduit expérimentalement sur le jeu de données mentionné dans cette introduction. Enfin, une conclusion au chapitre est proposée dans la section 3.4, avec quelques perspectives sur l'annotation sémantique des images et vidéos.

3.2 Extraction de mots-clés : état de l'art

L'extraction de mots-clés (*Keyword Extraction* ou KE) est définie par Beliga [12] comme le processus de sélection de l'ensemble des mots contenus dans un document qui décrivent le mieux ses thématiques. Une autre définition est proposée par Bharti et Babu [16] : l'extraction de segments-clés est le processus qui sélectionne les mots ou les phrases qui décrivent le mieux un texte, avec ou sans intervention humaine. Ainsi, en fonction des auteurs et des méthodes, les termes choisis ne se résument pas forcément à des mots-clés, mais peuvent aussi être des phrases-clés de taille variable. Pour généraliser, l'appellation segment-clé est donc souvent employée, un mot étant un segment particulier de taille 1.

L'extraction de mots-clés est une problématique à l'intersection des domaines de la fouille de texte (*Text Mining* ou TM), qui étudie l'extraction automatique de connaissances précédemment inconnues à partir de ressources écrites [67] ; de la recherche d'information (*Information Retrieval* ou IR), qui s'intéresse à l'accès optimal à l'information par l'indexation et l'interrogation [138] ; et du traitement du langage naturel ou TLN, qui cherche à rendre les langages naturels humains exploitables et compréhensibles par une machine [33]. De nombreuses applications de ces domaines font usage de l'extraction de mots-clés comme l'indexation, la représentation de documents pour classification [94, 119], la création de résumés (*Text Summarization*), la reconnaissance d'entités nommées (*Named Entity Recognition*) et la découverte de relations (*Relation Extraction*) [111], la construction de dictionnaires de domaine, la détection de thèmes, l'extraction automatique de mots-clés pour le référencement de sites web [168], etc.

À l'origine effectuée manuellement, le besoin de rendre l'extraction de mots-clés automatique (*Automatic Keyword Extraction* ou AKE) s'est fait ressentir à l'aube du Web puis lors de l'explosion des réseaux sociaux numériques, tous deux à l'origine d'un afflux massif de données, pour certaines non-structurées, aussi qualifié de *Big Data*. En effet, l'extraction manuelle était laborieuse et coûteuse en temps et en main d'œuvre, ce qui était incompatible avec l'augmentation du nombre de données et de documents à traiter. Il fallait alors trouver de nouvelles méthodes capables de passer à l'échelle, tout en gérant la dispersion de l'information et la difficulté de choisir les bons mots-clés.

Pour évaluer les résultats des méthodes, des jeux de données étiquetés sont utilisés. Ceux-ci contiennent un ensemble de textes accompagnés de la liste de leurs mots-clés, aussi qualifiés de mots-clés dorés ou en anglais *golden keywords*, attribués directement par les auteurs ou par les créateurs des jeux via *crowdsourcing*[2]. Quelques exemples de jeux de données étiquetés populaires en AKE sont présentés dans la table A.1 en annexe. À partir des mots-clés dorés, les mesures suivantes sont évaluées :

— les **vrais positifs** (TP), nombre de segments-clés sélectionnés qui correspondent à des mots-clés dorés ;

— les **faux positifs** (FP), nombre de segments-clés sélectionnés qui ne correspondent pas à des mots-clés dorés ;

— les **faux négatifs** (FN), nombre de mots-clés dorés non-détectés par la méthode ;

— la **précision**, qui donne une indication de pertinence en mesurant la proportion de résultats

2. Réquisition d'un très grand nombre d'individus pour manuellement faire un travail conséquent par partage de la charge.

qui sont réellement des mots-clés dorés ;

$$précision = \frac{TP}{TP + FP} \tag{3.1}$$

— le **rappel**, qui donne une indication de couverture en mesurant la proportion de mots-clés dorés qui ont été identifiés ;

$$rappel = \frac{TP}{TP + FN} \tag{3.2}$$

— le **F1-score**, qui fait la synthèse des informations précédentes grâce à leur moyenne harmonique.

$$f1 - score = 2 \times \frac{précision \times rappel}{précision + rappel} \tag{3.3}$$

Dans certains cas, quelques mesures supplémentaires comme la *binary preference measure* (Bpref) ou le *mean reciprocal rank* (MRR) chez Liu et al. [95] sont calculées, notamment lorsque l'ordre des segments-clés a une importance.

Dans la littérature, il n'existe pas de consensus sur la manière de déterminer la correspondance entre un segment-clé identifié par une méthode et un mot-clé doré. Les travaux de Turney [149] montrent que chercher une correspondance exacte (les deux doivent être strictement identiques) a pour conséquence de produire des résultats largement sous-estimés par rapport à des notes qui seraient attribuées par des juges humains. Comme mis en avant dans la revue de Hasan et Ng [65], ceci est dû à la souplesse du jugement humain par rapport aux variants d'extension ("artificial neural network"), morphologiques ("neural networks") et lexicaux ("neural net") d'un mot-clé de départ ("neural network"). Certaines méthodes font appel à des outils de TLN pour révéler les connexions entre les variants [14]. D'autres privilégient des correspondances partielles (l'un doit contenir l'autre), par exemple à travers certaines méthodes d'évaluation comme ROUGE [93], BLEU [121], METEOR [90] ou NIST [75]. Cette dernière approche ne fait néanmoins pas non plus l'unanimité [84]. Ces divergences dans les méthodes d'évaluation rendent difficile la comparaison de résultats annoncés dans des articles différents.

Cette difficulté est amplifiée par la dépendance de la qualité de l'extraction à certaines caractéristiques du texte. Certains facteurs sont réputés pour altérer l'efficacité des méthodes :

— le **type de document**, car certains formats ont un impact sur la disposition et la fréquence des mots-clés (ex : dans les articles scientifiques, les mots-clés sont concentrés dans l'abstract et dans l'introduction et un faible nombre de thématiques différentes non-corrélées sont abordées, contrairement aux articles journalistiques ou aux *blogs* [65]) ;

— la **taille des documents**, car l'éparpillement des segments-clés dans un grand espace de recherche peut être un obstacle sur les documents longs, tout comme la plus faible sémantique des documents trop courts [66] ;

— la **langue**, puisque certaines méthodes en sont dépendantes à cause de l'utilisation de techniques de TLN notamment.

Ainsi, une même méthode très efficace sur un ensemble de documents peut obtenir des valeurs de précision, rappel ou f1-score très mauvaises sur un autre.

Au cours du temps et en fonction des visions des différents auteurs, plusieurs classifications des approches d'AKE ont été proposées. En 2008, Zhang et al. [165] proposent 4 catégories de méthodes : les approches statistiques, linguistiques, *machine learning* et hybrides. Cette

vision, souvent reprise [16], est parfois simplifiée en 2 catégories : approches statistiques et approches *machine learning* [32]. En 2013, Menaka et Radha [106] font apparaître pour la première fois deux nouvelles familles d'approches qui gagnent en influence à cette époque, les approches basées graphes et celles basées réseaux de neurones. La classification proposée est alors : approches basées chaînes lexicales (anciennement linguistiques), approches basées co-occurrences (anciennement statistiques), approches basées graphes et approches basées réseaux de neurones. Avec la multiplication des catégories, les revues plus récentes comme celle de Beliga [12] ou celle de Nasar, Jaffry et Malik [112] optent toutefois plus pour une classification plus générale, basée sur deux catégories principales : les approches supervisées et les approches non-supervisées. À celles-ci, la première revue ajoute les approches semi-supervisées, là où la seconde rajoute plutôt les approches heuristiques et celles basées réseaux de neurones.

Dans cet état de l'art, j'ai décidé de suivre une version simplifiée de la classification proposée par Nasar, Jaffry et Malik [112]. En effet, après lecture de la revue, il ne me semble pas que les approches heuristiques soient une catégorie pertinente, dans le sens où les méthodes présentées par les auteurs sont toutes des approches supervisées ou non-supervisées, améliorées avec des règles basées sur la linguistique et/ou des statistiques. La classification que j'ai retenue est alors : **approches non-supervisées**, **approches supervisées** et **approches réseaux de neurones**. Chaque catégorie principale se décline ensuite en sous-catégories, induites par le type d'algorithme ou de représentation des données utilisé par les différentes méthodes. J'ai décidé de traiter les réseaux de neurones à part des autres méthodes d'apprentissage car, selon les cas, ces approches peuvent être supervisées ou non. Les traiter en tant que sous-parties des deux autres catégories principales aurait été redondant et n'aurait pas permis de bien comprendre l'évolution des usages au cours du temps et selon les besoins et limites des situations.

La figure A.4 jointe en annexe présente quelques statistiques sur la répartition des différentes approches au cours du temps, que j'ai calculées en étudiant les dates de publication des méthodes présentées dans cet état de l'art. Chronologiquement, les publications s'étendent de 1957 à 2020. On remarque un plus grand nombre de méthodes non-supervisées que supervisées. On observe également deux pics en 2010 et 2018, qui montrent que l'extraction automatique de mots-clés reste une problématique d'actualité. On note enfin une accélération de l'utilisation de méthodes basées réseaux de neurones ces dernières années, qui révèle peut-être un potentiel de cette famille d'approches par rapport aux autres, bien établies depuis des dizaines d'années.

Du côté des méthodes non-supervisées, la figure A.5 fait apparaître deux phases principales : la domination des méthodes statistiques (STA), possiblement améliorées par des heuristiques linguistiques (LIN), puis l'arrivée et la nouvelle domination des méthodes basées graphes (GRA). On remarque quand même que des approches statistiques continuent d'être développées plus récemment, ce qui indique une forme d'efficacité de leur part.

Sur la figure A.6, on observe des phases dans l'utilisation des algorithmes d'apprentissage supervisé. Les premières méthodes se sont tournées vers la classification naïve bayésienne (BAY), qui a ensuite laissé sa place aux arbres de décision (ARB), qui ont eux-mêmes laissé la leur aux méthodes SVM (*Support Vector Machine*). Après un âge d'or dans les années 2000, on peut observer un déclin de ces approches au début des années 2010, qui s'est résulté par le besoin d'étudier de nouvelles méthodes basées sur d'autres algorithmes de *Machine Learning* (AML).

La suite de cet état de l'art présente les différentes approches d'extraction automatique de mots-clés, leurs sous-catégories ainsi que leurs méthodes les plus importantes.

3.2.1 Approches non-supervisées

Les méthodes d'extraction de mots-clés non-supervisées sont présentées dans cette sous-section. Ces techniques sont caractérisées par l'absence de superviseur ou de modèle pour classifier les données. De fait, aucun jeu de données étiqueté n'est nécessaire à l'exécution de la méthode (si ce n'est éventuellement pour la validation des résultats a posteriori). En extraction automatique de mots-clés, les méthodes non-supervisées sont réparties dans deux sous-catégories principales :

— les **approches statistiques**, qui identifient les segments-clés selon leur importance dans le texte ;

— les **approches graphes**, qui capturent les relations sémantiques entre les mots pour identifier les segments-clés selon leur rôle au sein du texte.

Parfois, ces approches sont enrichies par des connaissances externes linguistiques.

3.2.1.1 Approches statistiques

Premières approches historiquement mises en œuvre pour identifier les mots-clés dans un texte, les approches statistiques cherchent à évaluer l'importance d'un mot par rapport aux autres en lui attribuant un poids, et à ensuite sélectionner les candidats les plus importants comme mots ou segments-clés. Les premiers travaux remontent à 1957 avec Luhn [98]. L'hypothèse simple formulée par l'auteur était que plus un terme est répété un grand nombre de fois, plus il a de l'importance dans un texte. La simple fréquence d'occurrence d'un mot était alors utilisée pour pondérer les termes.

Quelques années plus tard, en 1972, Karen Spärck Jones déplore la faiblesse de la fréquence simple comme pondération, notamment face aux mots communs, aussi appelés mots vides (*stop words*), comme les prépositions ou les verbes auxiliaires. Elle propose alors TF-IDF [81], pour Term Frequency-Inverse Document Frequency, une mesure qui sera par la suite très utilisée en AKE grâce à sa capacité à évaluer à la fois l'importance et la spécificité d'un mot. Pour cela, sa formule (équation 3.3) prend en considération sa fréquence dans le texte en cours d'analyse et sa fréquence d'utilisation dans d'autres textes d'un corpus. Ainsi, cette mesure permet de faire la différence entre un terme qui est régulièrement fréquent et qui pourrait alors être apparenté à un mot vide et un mot qui est fréquent car il représente la thématique d'un texte particulier. Jones montre dans son article que cette simple amélioration permet d'augmenter de manière significative les performances (précision, rappel, f1-score) de l'extraction de mots-clés. Toutefois, la mesure est inadaptée aux corpus de documents qui portent tous sur les mêmes thématiques. En effet, les termes relatifs à ce thème commun sont dans ce cas très régulièrement utilisés, et finissent par paraître comme des mots vides.

$$TFIDF_{i,j} = tf_{i,j} \times \frac{|D|}{|\{d_j : t_i \in d_j|} \tag{3.4}$$

Figure 3.3 – TF-IDF du terme t_i du document textuel d_j, avec $tf_{i,j}$ la fréquence de t_i dans d_i, $|D|$ le nombre d'autres textes du corpus et $|\{d_j : t_i \in d_j|$ le nombre de documents où le terme t_i apparaît.

La mesure est reprise en 1975 par Salton, Wong et Yang [135]. Dans cet article, les auteurs proposent d'utiliser une représentation vectorielle des mots pour pouvoir réduire la dimension d'un texte en le plaçant dans un espace euclidien. Ce modèle, appelé Vector Space Model ou VSM, utilise la valeur TF-IDF comme l'une des trois dimensions de l'espace. Les mots sont regroupés par un algorithme non-supervisé de type *k-means* [48] et les termes dont les coordonnées se rapprochent le plus des centroïdes des groupes sont sélectionnés comme mots-clés. Le principe de représentation vectorielle des textes sera ensuite beaucoup repris dans les autres approches d'extraction de mots-clés et de traitement du langage naturel de manière générale, notamment pour pouvoir alimenter les entrées des algorithmes d'apprentissage supervisé. Toutefois, le modèle VSM comme proposé ici possède trois limites importantes :

— la structure et la sémantique du texte ne sont pas du tout pris en considération ;

— chaque mot est considéré comme complètement indépendant des autres ;

— si deux textes développent la même idée avec des termes différents, la similarité ne peut pas être identifiée facilement.

Plutôt que d'utiliser la fréquence des mots, certains auteurs s'intéressent à la fréquence des co-occurrences, aussi appelés n-grammes. La valeur de n dans le terme précédent représente la taille de la co-occurrence : 1-gramme ou unigramme = occurrence de mot ; 2-gramme ou bigramme = co-occurrence de 2 mots ; 3-gramme ou trigramme = co-occurrence de 3 mots, etc. En plus de mesurer l'importance des termes d'un texte, cette approche permet également de considérer les relations sémantiques et donc d'identifier certains segments-clés par association.

La méthode de Cohen [36] privilégie les co-occurrences de caractères plutôt que de mots, ce qui garantit une indépendance à la langue et au domaine. Les segments-clés sont les mots ou séquences de mots qui contiennent les n-grammes de caractères les plus fréquents. Les méthodes de Steier et Belew [140], HaCohen-Kerner [62] et Matsuo et Ishizuka [101] s'intéressent aux co-occurrences de mots. La première utilise une estimation de l'information mutuelle portée par une co-occurrence à 2 termes, calculée à partir des fréquences normalisées de la co-occurrence et des termes individuels. La deuxième calcule la fréquence de tous les unigrammes, bigrammes et trigrammes et sélectionne simplement les n-grammes les plus fréquents. La dernière méthode sélectionne les mots les plus fréquents et les bigrammes dont ils font partie comme mots-clés. Dans ces trois articles, les valeurs de précision, rappel et f1-score obtenues expérimentalement surpassent celles de TF-IDF sur les mêmes jeux de données.

Une autre information régulièrement exploitée pour identifier les mots-clés est la position des termes. Certaines approches font par exemple l'hypothèse que les mots importants ont tendance à se concentrer dans certaines zones du texte, là où les mots vides et autres termes moins importants sont distribués de manière plus uniforme ou aléatoire. L'entropie de Shannon de la distribution des différentes occurrences d'un terme est étudiée à la fois par Herrera et Pury [68] et par Yang et al. [162]. Les termes avec une forte valeur d'entropie, qui sont donc plus uniformément répartis dans le texte, correspondent à des mots vides, là où les termes avec une faible valeur correspondent à des mots-clés. Une méthode inspirée du concept d'analyse statistique des systèmes quantiques désordonnés est proposée par Carpena et al. [30]. Les auteurs prennent à la fois en considération la fréquence des termes et l'écart-type des distances entre leurs occurrences, ce qui permet d'évaluer si les mots sont répartis aléatoirement ou non.

Parfois, seule la position de la première occurrence d'un mot ou d'un n-gramme est utilisée, comme dans la méthode KP-Miner de El-Beltagy et Rafea [14]. Dans leur article, les auteurs font deux observations importantes :

— plus un terme est important, plus il est susceptible d'apparaître "tôt" dans un texte ;

— au-delà d'un seuil de position donné, la probabilité pour un mot jamais apparu auparavant d'être un mot-clé chute de manière importante.

La méthode pondère donc les mots et n-grammes à partir de leur valeur TF-IDF et par rapport à la position de leur première occurrence. KP-Miner obtient de meilleurs résultats que KEA [158], la méthode de référence en extraction automatique de mots-clés supervisée, sur le jeu de données Turney (2000) (voir table A.1 en annexe) et sur un jeu de données construit à partir d'articles Wikipédia en arabe, ce qui appuie son indépendance par rapport à la langue.

Une autre approche non-supervisée basée statistiques, complètement indépendante du langage et du domaine, est proposée par Campos et al. [28] avec YAKE !. Les poids des mots du texte sont obtenus par combinaison de plusieurs caractéristiques liées à leur fréquence (fréquence moyenne, fréquence d'apparition dans des phrases différentes), à la position de leur première occurrence, à leur casse (présence de majuscules) et à leur probabilité d'être un mot vide. Tous les unigrammes, bigrammes et trigrammes sont considérés puis triés par poids décroissants. Les mots-clés séparés par une distance de Levenshtein inférieure à 0.8 sont considérés comme doublons et seul un des deux est conservé pour ne pas polluer la liste des résultats. Sur 4 jeux de données différents, les auteurs de YAKE ! réussissent à obtenir de meilleurs résultats que TF-IDF et que trois autres méthodes non-supervisées orientées graphes : TextRank [108], RAKE [132] et SingleRank [152].

3.2.1.2 Quelques mots sur les approches linguistiques

Les méthodes non-supervisées d'extraction de mots-clés basées sur la linguistique proposent d'utiliser des connaissances heuristiques externes sur le langage naturel pour mieux caractériser les termes importants. Traditionnellement, la littérature considère trois types d'informations linguistiques pouvant être utilisées pour l'extraction : les informations lexicales, qui portent sur les mots ; les informations syntaxiques, qui portent sur les phrases ; les informations sur le discours, qui portent sur la sémantique du texte.

Des informations sur le discours sont exploitées par Salton et Buckley [134], qui utilisent une encyclopédie pour améliorer les performances des méthodes statistiques en filtrant les mots-clés qui ne sont pas reliés à un titre d'article et en proposant une nouvelle mesure basée sur le nombre de références croisées entre les termes fréquents. Cette prise en compte de la sémantique permet d'améliorer au minimum de 50% la précision des méthodes de base, et de plus de 5% le rappel. Ceci se fait toutefois au prix de l'indépendance à la langue, comme très souvent avec les approches linguistiques, et l'exhaustivité et la qualité de l'encyclopédie choisie a une très grande importance. Une approche similaire mais basée sur un recueil d'articles journalistiques est développée par Dennis [39].

Des informations syntaxiques sont exploitées par Krulwich et Burkey [87], qui accordent un poids plus important aux mots qui possèdent un style particulier (italique, gras, etc.), aux termes utilisés dans les en-têtes (de page, de colonne, etc.) et à ceux qui sont ou qui possèdent des acronymes qui les suivent ou les précèdent. Des heuristiques fondées sur les balises des documents semi-structurés comme des pages HTML sont parfois proposées, par exemple par Humphreys [72] dans sa méthode PhraseRate, ou encore par Thomas, Bharti et Babu [146]. Dans ce cas, les balises HTML et leurs métadonnées permettent d'accorder une pondération adaptée à certains termes, notamment à ceux qui sont visuellement mis en avant sur les pages. Les informations

syntaxiques permettent d'appréhender le contexte d'une phrase, mais sont toutefois très sensibles au bon respect de la syntaxe par les auteurs des textes (correcte utilisation des styles, etc.).

Des informations lexicales permettent à Barzilay et Elhadad [10] d'améliorer les performances de la méthode basée sur la fréquence des n-grammes de HaCohen-Kerner [62], en nettoyant la liste de segments-clés candidats avec un outil populaire en TLN appelé *part-of-speech tagger* ou *POS tagger*, qui permet d'attribuer à chaque mot d'un texte une étiquette représentant sa nature grammaticale (nom, adjectif, verbe, etc.). Seuls les n-grammes qui correspondent à une séquence de natures autorisée (adjectif + nom, etc.) sont conservés. Les mots vides, identifiés à partir d'une liste prédéfinie, sont supprimés, puis les résultats obtenus passent par une dernière étape de racinisation (en anglais *stemming*), dont l'objectif est de retirer les préfixes et suffixes des termes pour ne garder que leur racine et ainsi supprimer les variants morphologiques (ex : "vaccin", "vaccination" et "vaccins"). Ces trois étapes successives de préparation des candidats est souvent reprise par d'autres méthodes de la littérature. Bien que très efficaces, leur impact positif sur les résultats dépend cependant beaucoup de la langue et des outils de TLN utilisés.

Enfin, certaines méthodes combinent l'utilisation d'informations linguistiques de différents types, par exemple pour construire des chaînes lexicales [110] comme Barzilay et Elhadad [10]. Dans cette méthode, les n-grammes du texte sont filtrés selon leur composition grammaticale, puis selon leur appartenance au dictionnaire électronique WordNet. Une pondération des chaînes lexicales est calculée grâce à une analyse statistique de la fréquence et de la position des mots et permet de sélectionner les candidats les plus importants. Selon une évaluation par des juges humains, la méthode propose des résultats plus pertinents qu'une approche statistique simple à base de n-grammes. Cependant, les auteurs remarquent que les phrases longues sont privilégiées.

Les états de l'art de nombreux articles sur l'extraction automatique de mots-clés considèrent les approches linguistiques comme une classe de méthodes à part entière. En ce qui me concerne, je ne suis pas d'accord avec cette représentation, puisque les approches linguistiques ne sont jamais utilisées directement pour discriminer les mots-clés. Elles servent plutôt à améliorer la précision, le rappel ou le f1-score des méthodes, et sont donc uniquement pour moi des techniques de nettoyage des données.

3.2.1.3 Approches graphes

Les méthodes d'extraction non-supervisées les plus efficaces aujourd'hui utilisent des graphes pour représenter le texte, puis considèrent les sommets les plus centraux comme mots-clés. Un graphe est constitué de sommets, qui en général ici sont les mots du texte, et d'arêtes, qui représentent les relations sémantiques entre les mots. En fonction des méthodes, le critère pour relier deux sommets par une arête varie, pouvant aller de la simple co-occurrence à un lien sémantique extrait d'une base de connaissances externe. Les arêtes des graphes sont généralement pondérées pour représenter la force de la connexion, par exemple avec la fréquence de la co-occurrence des mots aux extrémités. Un graphe peut être représenté par une matrice carrée dite d'adjacence, où la connexion entre chaque paire possible de sommets est représentée par son poids (0 si non-reliés, 1 par défaut).

La première méthode orientée graphe a été proposée par Ohsawa, Benson et Yachida [117], avec pour objectif de proposer un algorithme simple et rapide, applicable sur des textes seuls (sans corpus complet) et qui fait la distinction entre l'importance d'un terme et sa fréquence. Pour cela, la méthode, appelée KeyGraph, construit un graphe de co-occurrences par phrase.

En d'autres termes, les sommets sont des unigrammes, bigrammes ou trigrammes et sont reliés par une arête s'ils apparaissent dans la même phrase (les mots vides sont préalablement retirés). La pondération correspond à la fréquence de la co-occurrence. Des *clusters* sont construits à l'intérieur du graphe et des scores basés sur la fréquence et le degré (nombre de voisins) sont attribués aux sommets. Un certain nombre de segments-clés par *cluster* sont sélectionnés parmi les plus hauts scores. La méthode KeyWorld de Matsuo, Ohsawa et Ishizuka [102] utilise le même type de graphe, mais se passe de la construction des *clusters*. L'importance des sommets est évaluée en mesurant leur contribution à la propriété du petit monde, par l'étude des plus courts chemins où ils sont impliqués, et est pondérée avec leur valeur IDF. Dans les deux articles, des évaluations sur des articles scientifiques révèlent une augmentation significative de la valeur de rappel par rapport à TF-IDF, liée à une meilleure détection des segments-clés peu fréquents.

Une autre utilisation des graphes est faite par Erkan et Radev [43] dans LexRank. Une matrice de similarité cosinus entre paires de segments-clés candidats est construite et sert ensuite de matrice d'adjacence à un graphe dit de similarité. Le score attribué pour départager les termes est leur centralité de vecteur propre [19]. Ce principe est également utilisé par la méthode SemCluster de Alrehamy et Walker [4], mais avec comme mesure de similarité WuPalmer. Les segments-clés ne sont pas sélectionnés selon une valeur de centralité mais grâce à des règles heuristiques, appliquées au sein de *clusters* construits par propagation d'affinité. La méthode propose également une phase de préparation où les candidats sont filtrés selon leur composition grammaticale grâce à un *POS tagger* et sont sémantiquement désambiguïsés grâce à un autre algorithme de TLN. Une dernière méthode similaire est proposée par Palshikar [119]. Une mesure de dissimilarité est utilisée pour construire la matrice d'adjacence, et seules les arêtes reliant des termes qui co-occurrent au moins une fois dans la même phrase sont conservées. Les résultats expérimentaux obtenus avec ces méthodes surpassent tous une nouvelle fois ceux de TF-IDF sur les mêmes jeux de données.

En 2004, Mihalcea et Tarau [108] proposent TextRank, une méthode basée graphes qui est aujourd'hui encore très utilisée comme point de comparaison avec d'autres méthodes d'extraction de mots-clés. Après nettoyage du texte par suppression des mots vides et filtrage par analyse des natures grammaticales, une fenêtre de 2 mots le parcourt. Chaque co-occurrence amène à la création de deux sommets reliés par une arête non-dirigée, sauf si les termes sont séparés par une ponctuation. Dans ce graphe de co-occurrences construit par fenêtre glissante, les arêtes sont pondérées par la fréquence du bigramme et les sommets sont départagés par leur centralité PageRank [26], une mesure de popularité qui prend en considération la popularité des voisins. Les résultats obtenus lors des évaluations expérimentales par correspondance partielle (méthode ROUGE) sur des articles journalistiques surpassent ceux de la méthode supervisée de Hulth [71], à part pour la valeur de rappel.

De nombreuses méthodes ont par la suite étendu TextRank. Islam et Islam [74], avec Graph Based Random Walk Model (GBRWM), proposent de pondérer différemment les arêtes en considérant les valeurs TF-IDF des sommets aux extrémités. De même pour Pan, Li et Dai [120], qui ajoutent en plus l'entropie moyenne de l'information portée par la co-occurrence au calcul de la pondération. La méthode DegExt proposée par Litvak et al. [94] reprend le même principe de la fenêtre glissante de taille 2 que TextRank, mais utilise des arêtes dirigées et remplace la centralité PageRank par la centralité de degré. Les arêtes ne sont pas pondérées mais étiquetées avec les identifiants des phrases qui contiennent la co-occurrence, dans le but de faciliter la construction de phrases-clés en suivant les plus courts chemins avec les plus grandes centralités moyennes. Selon le comparatif effectué par les auteurs de DegExt sur un jeu de données constitué

d'articles journalistiques en anglais, cette approche améliore de 15% la précision de TextRank sur les documents longs, au prix de plus faibles valeurs de rappel et de f1-score.

Une extension très populaire de TextRank est SingleRank, proposée par Wan et Xiao [152]. La taille de la fenêtre glissante est augmentée à 10 mots afin de mieux considérer le contexte des termes. Comme dans DegExt, les n-grammes qui possèdent les sommes de centralité les plus élevées sont sélectionnés comme segments-clés supplémentaires. Une variation de cette méthode est proposée dans le même article, ExpandRank. Contrairement à la première méthode proposée, qui se voulait applicable sur des documents isolés, cette deuxième approche exploite des informations tirées des autres textes d'un corpus comme peut le faire TF-IDF. Une mesure de similarité cosinus est calculée entre tous les textes du corpus. D'un côté, la méthode TextRank est appliquée de manière traditionnelle sur le texte, et d'un autre, elle l'est sur un graphe global construit à partir de la fusion des textes les plus similaires à celui analysé. Les valeurs de centralité finales sont obtenues en prenant en compte les valeurs locales et globales. ExpandRank permet d'obtenir de meilleures performances que SingleRank, mais est plus gourmande en temps d'exécution et nécessite un corpus de plusieurs textes. Les deux méthodes permettent d'obtenir de meilleurs résultats que TextRank.

La méthode SingleRank est elle-même étendue par plusieurs autres auteurs. La moyenne harmonique plutôt que la somme pour la construction des n-grammes est proposée par Yeom, Ko et Seo [163] afin de ne pas favoriser les termes longs. Florescu et Caragea [46], quant à eux, proposent avec PositionRank de modifier le calcul de la centralité PageRank pour qu'il prenne en considération la distribution des occurrences des termes ainsi que la position de leur première occurrence. Sur tous les jeux de données testés, les performances de PositionRank s'avèrent meilleures que celles de SingleRank et ExpandRank. De fait, cette méthode est très régulièrement citée comme méthode de pointe (*state-of-the-art*) dans la littérature.

Si plusieurs sujets sont abordés dans un même texte, TextRank ne permet pas de garantir qu'ils soient tous au moins représentés par un segment-clé dans les résultats. Pour surmonter cette limite, une allocation de Dirichlet latente (LDA) [17] est utilisée dans la méthode Topical PageRank de Liu et al. [95]. Elle permet de détecter les différents sujets abordés dans un texte et d'attribuer aux sommets une probabilité d'appartenir à chaque thématique. Autant de graphes que de sujets détectés sont construits de la même manière que dans TextRank, en ne prenant pas en considération les mots avec une probabilité nulle. Les scores finaux pour départager les sommets sont obtenus en calculant la somme de toutes les centralités PageRank, pondérée avec les probabilités d'appartenir aux différents sujets. Avec Topical PageRank, les auteurs parviennent à obtenir de meilleures valeurs de précision, rappel et f1-score qu'avec TF-IDF et TextRank.

Le principe de Topical PageRank a été étendu de nombreuses fois par la suite. Pour identifier les thématiques principales du texte, Grineva, Grinev et Lizorkin [57] utilisent un algorithme de détection de communautés [114] plutôt qu'une LDA dans CommunityCluster. Un groupement hiérarchique agglomératif est exploité dans TopicRank de Bougouin, Boudin et Daille [24] pour regrouper les mots par sujet selon leur similarité de Jaccard. TopicCoRank [23] est une extension de la méthode précédente par les mêmes auteurs, qui rajoute une surcouche d'affectation de mots-clés en plus de l'extraction à partir d'une ontologie de domaine. Une amélioration du temps d'exécution de Topical PageRank, sans altérer ses performances, est obtenue par Sterckx et al. [141] dans Single Topical PageRank. Plutôt que de relancer TextRank dans chaque sous-graphe, les résultats sont déduits à partir de la similarité cosinus entre les sujets détectés par la LDA. Enfin, une nouvelle mesure complémentaire à la centralité PageRank, la *salience* (protubérance), est considérée dans la méthode SalienceRank de Teneva et Cheng

[145] pour évaluer la spécificité des termes. Cette valeur est une combinaison linéaire entre la spécificité du terme dans le sujet (à quel point le mot est partagé par plusieurs thèmes) et sa spécificité dans le corpus de textes (IDF).

Bien que TextRank ait popularisé l'utilisation de la centralité PageRank pour déterminer l'importance des termes, certaines méthodes ont pu étudier l'apport d'autres mesures. Ainsi, Huang et al. [70] reprennent le principe de TextRank mais utilisent la centralité d'intermédiarité, qui estime le nombre de fois où un sommet se trouve sur le plus court chemin entre deux sommets quelconques du graphe pour déterminer son influence sur le transfert de l'information. Une comparaison des différentes mesures de centralité sur des jeux de données en anglais et en français est proposée par Boudin [22]. Les résultats obtenus montrent que sur des documents longs (articles scientifiques), la centralité de degré permet d'obtenir d'aussi bonnes performances que TextRank. Sur des documents courts, en revanche, la centralité de proximité est celle qui identifie le mieux les segments-clés. Cette dernière observation est reprise et confirmée par Abilhoa et De Castro [1] sur un jeu de données constitué de *tweets*, dont la taille maximale autorisée était de 140 caractères à l'époque de l'étude. Un comparatif similaire avec d'autres mesures supplémentaires comme la force d'un sommet (degré pondéré par le poids de ses arêtes incidentes) ou son indice de dégénérescence [92] est menée par Lahiri, Choudhury et Caragea [88] sur 4 jeux de données différents. Les résultats obtenus sont comparables à ceux de Boudin. Les auteurs montrent que toutes les centralités permettent d'obtenir de meilleurs résultats que TF-IDF et que le degré et la force possèdent le meilleur rapport pertinence-temps d'exécution. Sur des documents courts, Tixier, Malliaros et Vazirgiannis [147] obtiennent de meilleurs résultats avec l'indice de dégénérescence [92], qui considère plus la cohésion des termes que leur popularité, qu'avec la centralité PageRank.

De nouvelles mesures de centralité sont également proposées pour identifier les segments-clés. La sélectivité est évaluée par Beliga, Meštrović et Martinčić-Ipšić [13] en tant que quotient du degré et de la force d'un sommet. Sur un jeu de données constitué d'articles journalistiques en croate, transformés en graphes en suivant la méthodologie proposée par TextRank, les auteurs obtiennent de meilleures valeurs de précision, rappel et f1-score qu'avec les centralités de degré, d'intermédiarité et de proximité. Leur méthode s'avère également peu sensible aux mots vides. Les créateurs de RAKE, Rose et al. [132], qui utilisent une mesure identique à la sélectivité, obtiennent même de meilleurs résultats que TextRank sur des documents courts. Après avoir constaté que, sur les graphes construits en suivant la méthodologie de TextRank, toutes les mesures de centralité sont corrélées à différents niveaux, Vega-Oliveros et al. [151] proposent MCI. Il s'agit d'une combinaison des centralités de degré et de vecteur propre, qui présentent la plus faible corrélation, ainsi que d'une mesure des trous structuraux [27]. Sur des documents courts, MCI départage mieux les candidats que les centralités de degré, d'intermédiarité, de proximité, PageRank, HITS et de vecteur propre.

Enfin, une dernière catégorie de méthodes basées graphes utilisent des modèles de plongement de mots ou de phrases, qui permettent d'apprendre une représentation des termes sous forme de vecteurs de nombres réels. Bien que le plongement utilise des modèles entraînés lors de phases d'apprentissage supervisées, j'ai décidé d'en parler dans cette partie et non pas dans la suivante car l'apprentissage ne concerne pas l'extraction des mots-clés en elle-même mais l'apport d'informations supplémentaires sur le texte, qui sont ensuite utilisées dans une méthode non-supervisée. Autrement dit, la sortie seule des algorithmes d'apprentissage supervisé utilisés dans les méthodes suivantes ne permet pas d'indiquer si un terme est un mot-clé ou non. Dans ce sens, le plongement est utilisé comme un outil, au même titre qu'un *POS tagger* par exemple,

et une version pré-entraînée sur d'autres jeux de données peut être exploitée.

Les auteurs Wang, Liu et McDonald [154] proposent de modifier la valeur de pondération des arêtes de TextRank par une combinaison de la fréquence des termes aux extrémités, de celle de leur co-occurrence et de la distance cosinus entre leurs représentations vectorielles, obtenues par un plongement de mots pré-entraîné sur des articles de Wikipédia. Une extension de la méthode précédente est proposée par Mahata et al. [100]. En plus d'agrandir la taille de la fenêtre glissante à 5 mots, une hypothèse est faite sur le texte : sa ou ses premières phrases synthétisent la sémantique des suivantes (sur des articles scientifiques, qui commencent par un *abstract*). Ainsi, un vecteur initial est calculé en sommant les représentations vectorielles des premiers mots. Les distances cosinus entre les vecteurs des deux termes aux extrémités et le vecteur initial sont également ajoutées à la formule de la pondération. Enfin, Ramiandrisoa [128] propose une méthode qui fait la synthèse des articles de Boudin [22], Wang, Liu et McDonald [154] et Yeom, Ko et Seo [163]. Une fois de plus, un graphe est exploité. Les sommets correspondent aux adjectifs et noms du texte, les arêtes sont non-dirigées et représentent la co-occurrence de deux termes dans une fenêtre glissante de taille 2, la pondération est le produit de la fréquence du bigramme et de la distance cosinus des représentations vectorielles des mots aux extrémités, les mots-clés sont les sommets qui possèdent les plus grandes centralités de degré et des phrases-clés sont construites à partir des plus courts chemins du graphe qui possèdent les plus grandes moyennes harmoniques de centralité et qui respectent les séquences de natures grammaticales autorisées. Sur des documents longs, cette approche surpasse TextRank, SingleRank, PositionRank et Topical PageRank.

3.2.1.4 Autres méthodes non-supervisées

Quelques méthodes exploitent des *clusters* de termes sans passer par la construction d'un graphe. L'attraction entre les mots, calculée à partir des écarts-types des distances moyennes entre les occurrences, est utilisée par Ortuño et al. [118] pour construire des groupes de termes. Le *cluster* avec l'attraction la plus forte est constitué des mots-clés du texte. L'approche KeyCluster de Liu et al. [96] utilise Wikipédia pour regrouper les mots par thématique, puis sélectionne les termes les plus proches des centroïdes de chaque *cluster* comme segments-clés. Dans la méthode de Pasquier [122], les mots sont regroupés puis départagés selon les probabilités d'appartenir aux différents sujets du texte attribuées par une LDA.

Les outils de plongement sont aussi exploités en dehors des approches basées graphes. Dans la méthode EmbedRank de Bennani-Smires et al. [15], un modèle de plongement de phrases et un autre de plongement de mots sont utilisés. Le texte est représenté par la somme des représentations vectorielles de ses phrases. Les mots qui possèdent les plus grandes valeurs de *Maximal Marginal Relevance* (MMR) [29] entre leur vecteur et celui du texte sont sélectionnés comme segments-clés. La méthode SIFRank de Sun et al. [142] utilise le même principe. Un modèle pré-entraîné de plongement de phrases SIF [7], qui est construit pour détecter et prendre en considération les thématiques du texte, est combiné à un modèle de plongement de mots pré-entraîné ELMo [124], qui est décrit par ses auteurs comme profondément contextualisé. Contrairement à la méthode précédente, le score utilisé pour départager les termes est la distance cosinus entre les vecteurs.

3.2.1.5 Conclusion sur les approches non-supervisées

Pour conclure, les méthodes d'extraction automatique de mots-clés non-supervisées se caractérisent par l'exploitation d'**informations** principalement **statistiques**, **structurelles** et/ou **linguistiques** dans le but d'attribuer à chaque mot du texte un **poids** qui départage les segments-clés candidats selon leur **importance** et/ou leur **spécificité**. Les approches non-supervisées ont l'avantage de ne pas nécessiter de **jeu d'entraînement étiqueté** et sont pour la plupart **indépendantes à la langue**. Cependant, les comparatifs avec des méthodes supervisées [108, 132] montrent que, même si certaines de ces méthodes parviennent à obtenir une **meilleure précision**, la valeur de **rappel** reste **toujours inférieure**. Ainsi, parmi les termes extraits par les méthodes non-supervisées, seule une faible proportion ne sont pas des mots-clés ou des parties de mots-clés (si correspondance partielle). Mais en même temps, seule une petite partie de l'ensemble des mots-clés à détecter le sont. L'utilisation d'**heuristiques** comme la fréquence ou la position peut être une explication à ces résultats. En effet, les termes fréquents sont majoritairement des mots-clés, d'où l'utilisation de cette mesure, mais tous les mots-clés ne sont pas fréquents, ce qui correspond à la réalité traduite par les valeurs de précision et de rappel. Selon moi, la **détection des mots-clés peu fréquents** est la clé pour améliorer les approches non-supervisées.

3.2.2 Approches supervisées

Les méthodes d'extraction de mots-clés supervisées sont présentées dans cette sous-section. On qualifie d'approches supervisées toutes les techniques basées sur la construction d'un modèle à l'aide d'un jeu de données d'entraînement étiqueté et d'une phase d'apprentissage pour ensuite généraliser l'expérience acquise pour extraire des mots-clés de nouveaux textes. L'extraction supervisée est généralement reformulée en problème de classification binaire, où chaque mot d'un texte se voit attribuer la classe "Keyword" ou "Not Keyword". Elle peut également parfois être reformulée en problème de classement par score (*ranking*) ou d'étiquetage de chaînes de caractères (*string labeling*).

3.2.2.1 Premières méthodes

Les deux premières méthodes supervisées datent de 1999 et sont aujourd'hui encore des points de comparaison incontournables pour les méthodes supervisées ou non.

GenEx de Turney [149] utilise la fréquence, la position, la taille et la nature grammaticale des mots pour construire des représentations vectorielles ensuite exploitées par un algorithme génétique de classification binaire. L'entraînement du modèle est effectué sur le jeu de données du nom de l'auteur référencé dans le tableau A.1 en annexe, c'est-à-dire sur un ensemble d'articles scientifiques, d'*e-mails* et de pages web. L'évaluation des résultats est faite manuellement par des juges humains, qui considèrent que 80% des segments-clés identifiés par GenEx sont de qualité acceptable. Deux limites principales sont identifiées par l'auteur : la dépendance aux outils de TLN (*POS tagger*) et donc au langage ; la faiblesse face aux synonymes, qui réduisent la fréquence des segments-clés.

La deuxième méthode supervisée historique est KEA de Witten et al. [158]. Chaque terme est de nouveau représenté par un vecteur, composé cette fois-ci de la valeur TF-IDF et de la position de la première occurrence du mot. Un modèle entraîné par apprentissage supervisé, ici un classifieur naïf bayésien, utilise ensuite cette représentation pour attribuer à chaque mot une

classe correspondant à son statut (segment-clé ou non). L'entraînement du modèle utilise 1300 articles scientifiques en anglais et 500 autres servent à l'évaluation par correspondance partielle. En moyenne, 2 mots-clés sur 5 sont correctement identifiés.

3.2.2.2 Classification naïve bayésienne

D'autres méthodes basées sur la classification naïve bayésienne étendent KEA. Les classifieurs de ce type utilisent un modèle probabiliste construit autour du théorème de Bayes, qui permet d'estimer la vraisemblance qu'un mot soit un segment-clé grâce aux probabilités [40, 131].

Le créateur de GenEx, Turney [148], remarque le manque de cohérence, notamment sémantique, entre les mots-clés identifiés par KEA, caractérisé par la présence de termes inattendus et hors propos. Il propose alors de rajouter une deuxième classification naïve, alimentée par les représentations vectorielles de KEA auxquelles sont ajoutées la probabilité d'être un mot-clé, attribuée lors du premier passage, et le rang du terme quand les candidats sont triés par probabilité décroissante.

D'autres auteurs proposent des extensions qui font varier les caractéristiques utilisées dans les représentations des mots. Des informations locales sur le contexte et sur les liens entre les mots sont ajoutées par Tang et al. [144] dans le but de supprimer l'indépendance des termes. Plus d'informations spatiales liées à la position des mots par rapport au début du document, du paragraphe et de la phrase sont exploitées par Uzun [150].

Des informations linguistiques sont également parfois prises en considération. KEA++ de Medelyan et Witten [105] ajoute aux caractéristiques de la méthode d'origine la taille du terme et le nombre de candidats auxquels il est sémantiquement lié, déterminé à l'aide d'un thésaurus. Des valeurs supplémentaires sont calculées par Nguyen et Kan [115] en fonction de la nature grammaticale du mot et de la présence d'acronymes ou de suffixes particuliers (ex : -ion ou -ment). Des booléens permettent également de décrire son appartenance à certaines sections particulières comme les titres, l'*abstract*, l'introduction, etc. Enfin, une extraction en profondeur de *patterns* de séquences de mots est utilisée par Xie, Wu et Zhu [160] pour obtenir 4 nouvelles caractéristiques qui capturent les relations sémantiques entre les mots. Toutes ces extensions permettent d'améliorer de manière significative les performances de KEA.

SurfKE de Florescu et Jin [47] est une solution supervisée qui exploite une classification naïve bayésienne mais qui ne se base pas sur KEA. Pour s'affranchir des connaissances de domaine nécessaires à l'identification des caractéristiques à utiliser, la méthode utilise un modèle de plongement de graphes basé sur une marche aléatoire sur un graphe de co-occurrences construit par fenêtre glissante (voir section 3.2.1.3). Les représentations vectorielles des sommets obtenues alimentent ensuite la classification. Une fusion des jeux Inspec et DUC 2001 (voir table A.1 en annexe) est utilisée pour l'entraînement et pour l'évaluation. La précision obtenue surpasse celle de KEA de plus de 50%. Les performances générales (précision, rappel et f1-score) dépassent largement celles des méthodes supervisées KEA et Maui [104], et non-supervisées KP-Miner [14] et PositionRank [46].

3.2.2.3 Arbres de décision

Pour la classification binaire, les arbres de décision sont également régulièrement exploités. Ces algorithmes représentent le processus de décision par un arbre, généré pendant la phase d'entraînement, où chaque feuille correspond à une sortie de l'algorithme et chaque nœud à un choix.

Lors de la présentation de GenEx, Turney [149] compare son algorithme génétique à une forêt d'arbres de décision C4.5 [126] à un vote, où les choix dans les arbres sont départagés par une mesure d'entropie et où les segments-clés sélectionnés sont ceux qui apparaissent dans le plus de listes générées par les modèles. Pour améliorer la pertinence des résultats, l'auteur conseille d'utiliser le méta-algorithme *Bagging* (*Bootstrap aggregating*), qui permet d'améliorer la robustesse et de réduire le surapprentissage (difficulté à généraliser) en appliquant une classification et une régression statistiques. Dans son étude, Hulth [71] confirme l'impact positif du *Bagging*. Les informations linguistiques, notamment sémantiques alimentées par des bases de connaissances externes, sont également identifiées comme le meilleur moyen d'augmenter la valeur de rappel des méthodes supervisées. Cette observation est encore plus vraie lorsque des arbres de décision sont utilisés, comme le montrent Krapivin et al. [86]. En effet, les mêmes connaissances linguistiques appliquées à un arbre de décision permettent d'obtenir un f1-score supérieur de 8% à celui obtenu par une classification naïve bayésienne, à scores équivalents sans.

Plusieurs méthodes de la littérature suivent le conseil de Turney [149] et utilisent des algorithmes C4.5 à base d'arbres de décision renforcés par *bagging* pour la classification. Medelyan, Frank et Witten [104] proposent Maui, qui exploite une représentation vectorielle composée de caractéristiques sémantiques, calculées grâce à la base de connaissances Wikipédia, statistiques (TF-IDF) et spatiales (position première occurrence). De même pour Lopez et Romary [97] avec HUMB, qui ajoutent toutefois quelques valeurs supplémentaires aux mêmes catégories comme la fréquence simple, la taille du terme, son *Generalized Dice Coefficient* (GDC) ou encore quelques valeurs booléennes pour indiquer la présence du terme dans une section importante comme le titre ou l'abstract. Une deuxième base de connaissances, GRISP, est également employée en complément. À méthodologie et représentations vectorielles identiques, les auteurs de HUMB montrent que les valeurs de précision, rappel et f1-scores obtenues sur le jeu de données SemEval (voir table A.1 en annexe) avec les arbres de décision renforcés par *bagging* surpassent celles obtenues avec l'algorithme SVM et sont équivalentes à celles proposées par un réseau de neurones, pour un temps d'exécution 57 fois moins important.

Des algorithmes C4.5 sans *bagging* sont utilisés par Song, Song et Hu [139] dans KPSpotter et par Ercan et Cicekli [42]. Ils adressent cette fois-ci le conseil de Hulth [71] concernant l'utilisation d'informations linguistiques. Dans la première méthode, une représentation vectorielle simple des mots par leur valeur TF-IDF et la position de leur première occurrence est utilisée, mais seuls les termes de certaines natures grammaticales sont considérés. Dans la seconde, un algorithme de construction de chaînes lexicales similaire à celui employé par la méthode non-supervisée de Barzilay et Elhadad [10] est employé pour filtrer les mots et phrases-clés candidats.

Enfin, d'autres algorithmes à base d'arbres de décision que C4.5 sont parfois utilisés, notamment son extension J48. Sur un jeu de données constitué d'articles scientifiques en anglais, HaCohen-Kerner, Gross et Masa [63] montrent que cette dernière méthode permet d'obtenir de meilleurs résultats qu'une sélection d'algorithmes de classification supervisés, dont la classification naïve bayésienne, SVM et C4.5. En utilisant des caractéristiques statistiques, spatiales et sémantiques obtenues par extraction de *patterns* de séquences de mots ainsi que l'algorithme de classification J48, Feng et al. [45] parviennent à améliorer de 50% la précision de KEA.

3.2.2.4 Support Vector Machines

Les séparateurs à vaste marge, ou en anglais *Support Vector Machines* (SVM) [60], sont également étudiés. Avec cette méthode, les éléments à classifier sont représentés par des vecteurs à l'intérieur d'un espace à plus ou moins grande dimension. Le modèle, composé d'autres vecteurs dits "supports", permet de séparer toutes les classes tout en maximisant la taille de la marge autour de la séparation.

Deux utilisations différentes des SVM sont possibles en AKE. Certains auteurs comme Zhang et al. [166] ou Hong et Zhen [69] avec Extended TF les utilisent pour effectuer une classification binaire. La première méthode, à destination des articles scientifiques, utilise une représentation vectorielle à base de caractéristiques globales (TF-IDF, position, nombre d'occurrences dans différentes parties) et locales (nature grammaticale, nombre de variants, somme TF-IDF des co-occurrences). La seconde méthode construit les vecteurs des termes selon certaines caractéristiques statistiques, spatiales et linguistiques (TF-IDF, position, nature grammaticale, etc.).

D'autres auteurs comme Jiang [77] ou Eichler et Neumann [41] avec DFKI KeyWE utilisent l'algorithme SVM, et plus précisément son extension SVM^{rank} [80], pour répondre à un problème de classement plutôt qu'à un problème de classification. Avec une représentation vectorielle des noms du texte à partir de leurs caractéristiques statistiques et spatiales, Jiang [77] obtiennent de meilleurs résultats qu'avec une classification binaire par SVM et qu'avec KEA (classification naïve bayésienne) sur des articles scientifiques et des pages web en chinois. La deuxième méthode suit le même principe et exploite aussi des caractéristiques statistiques et spatiales pour leurs représentations vectorielles, avec en plus quelques autres obtenues par l'étude de sources externes, notamment Wikipédia et un historique du moteur de recherche Bing.

3.2.2.5 Autres algorithmes de Machine Learning

Occasionnellement, d'autres algorithmes d'apprentissage supervisé sont utilisés pour étudier leur potentiel en AKE. Un modèle de Markov caché supervisé est utilisé par Conroy et O'leary [38] pour extraire les phrases-clés de rapports techniques et ensuite construire automatiquement des résumés. Une classification par régression logistique est proposée par Yih, Goodman et Carvalho [164] et par Haddoud et Abdeddaïm [64]. La première méthode est spécialisée pour les pages web. Un historique de moteur de recherche est utilisé en ressource externe et est couplé à de nouvelles caractéristiques de représentation des mots, calculées en exploitant la structure HTML (présence dans certaines balises ou méta-balises, textes d'hyperliens ou d'images, *parsing* d'URL, etc.). Les valeurs de précision, rappel et f1-score, évaluées de manière robuste grâce à une validation croisée, surpassent largement celles obtenues avec KEA et GenEx. La seconde proposition exploite des représentations vectorielles constituées de plus de 18 caractéristiques sur les mots, dont 6 nouvelles définies pour la méthode, et obtient de meilleurs résultats que HUMB, Maui ou KP-Miner sur le jeu de données SemEval 2010 (voir table A.1).

Le modèle des Champs Aléatoires Conditionnels (*Conditional Random Fields* ou CRF) est au centre de la méthode CRF++ de Zhang et al. [165]. L'extraction de mots-clés est considérée comme un problème d'étiquetage de chaînes de caractères (*string labeling*), comme le *POS tagging*. Les étiquettes attribuées représentent le statut du terme entre mot-clé, partie de phrase-clé, mot vide ou aucune de ces propositions. Les mots-clés sont ensuite identifiés par leur étiquette et les phrases-clés reconstruites selon certains *patterns*. Après entraînement et évaluation sur un jeu de 600 articles scientifiques en chinois, l'approche CRF obtient une précision jusqu'à 3 fois plus élevée que les approches statistiques simples basées sur TF-IDF et bien supérieure aux

approches supervisées SVM, régression logistique et régression linéaire multiple. Une extension de ce travail qui incorpore des connaissances sémantiques apportées par Wikipédia est proposée par Gollapalli, Li et Yang [54]. Cette modification permet aux auteurs de doubler les valeurs de précision, rappel et f1-score de KEA sur leur jeu d'articles scientifiques.

Enfin, l'approche non-supervisée basée graphes TextRank est étendue par Bordoloi et al. [20] avec Supervised Cumulative TextRank (KESCT) en remplaçant la pondération par des valeurs d'information mutuelle appelées *Unique Statistical Supervised Weights* (USSW) et calculées de manière supervisée. L'objectif d'une telle fusion des approches est, selon les auteurs, de pouvoir profiter de la prise en considération des relations sémantiques entre les mots et de la structure du texte apportée par les approches non-supervisées basées graphes, tout en bénéficiant des meilleures valeurs de rappel des méthodes supervisées.

3.2.2.6 Conclusion sur les approches supervisées

Pour conclure, les approches supervisées d'extraction de mots-clés se caractérisent par la construction d'un **modèle** à l'aide d'un **jeu de données annoté** et d'une **phase d'entraînement**, qui est ensuite utilisé pour différentes tâches comme de la **classification binaire**, du **classement** ou de l'**étiquetage de chaînes de caractères** pour obtenir une liste ordonnée ou non de segments-clés. Aucun algorithme d'apprentissage particulier ne semble faire consensus, même si les **arbres de décision** améliorés par ***bagging*** semblent obtenir les résultats les plus satisfaisants. La problématique principale de ces approches réside dans le **choix des caractéristiques** utilisées pour représenter les mots du texte étudié. Les **caractéristiques statistiques**, basées sur la fréquence d'apparition des termes (et ses variantes), et celles **spatiales**, basées sur la position des termes (première et dernière occurrence, distribution spatiale, etc.), sont majoritairement utilisées par les méthodes étudiées dans cette section. Plusieurs études montrent l'impact positif des **caractéristiques linguistiques**, calculées à partir d'outils de TLN comme les *POS tagger*, et des **caractéristiques sémantiques**, calculées à partir de ressources externes comme des bases de connaissances ou des thésaurus, mais cela bien souvent **au prix de l'indépendance à la langue**. L'utilisation de graphes pourrait cependant être une piste pour répondre à ce problème, puisqu'ils permettent à la fois de représenter la sémantique entre les termes et de révéler des caractéristiques linguistiques par la détection de *patterns*.

3.2.3 Approches basées réseaux de neurones

Les méthodes d'extraction basées sur des réseaux de neurones (RN) sont présentées dans cette sous-section. Ces approches peuvent être supervisées ou non selon les cas, mais se caractérisent toutes par l'utilisation d'un RN qui exploite une représentation vectorielle des mots, extraite automatiquement ou non, pour déterminer s'ils sont des segments-clés.

Un RN est un système probabiliste qui s'inspire du fonctionnement des neurones biologiques pour classifier des données en apprenant de leurs caractéristiques. Une version simplifiée et généralisée d'un réseau de neurones est présentée sur la figure 3.4. Ces systèmes s'apparentent à des graphes multipartites, en général dirigés de gauche à droite. Chaque neurone de la première couche, la couche d'entrée, correspond à une caractéristique de l'élément à classifier et donc à une dimension de sa représentation vectorielle. Chaque neurone de la dernière couche, la couche de sortie, correspond à une classe attribuable, qui peut être binaire (ex : mot-clé/pas mot-clé) ou à valeurs multiples (ex : mot-clé/pas mot-clé/partie de phrase-clé). Entre les deux, un certain nombre de couches dites cachées, de taille variable, peuvent être insérées. Les liens

entre les couches, appelés synapses, sont pondérés par des probabilités et chaque neurone est associé à une valeur réelle, calculée par une fonction de combinaison. Lorsque les poids des synapses ne sont pas attribués manuellement mais automatiquement par expérience, les RN sont qualifiés de perceptrons. Dans ce cas, les valeurs de pondération sont affinées pendant la phase d'apprentissage par rétropropagation. C'est-à-dire que chaque neurone adapte les valeurs de ses synapses en entrée en évaluant l'importance qu'ils ont eue sur l'erreur propagée dans les sorties, habituellement mesurée par le gradient de l'erreur. Plusieurs variantes de ce modèle de base existent et sont exploitées dans les méthodes qui suivent.

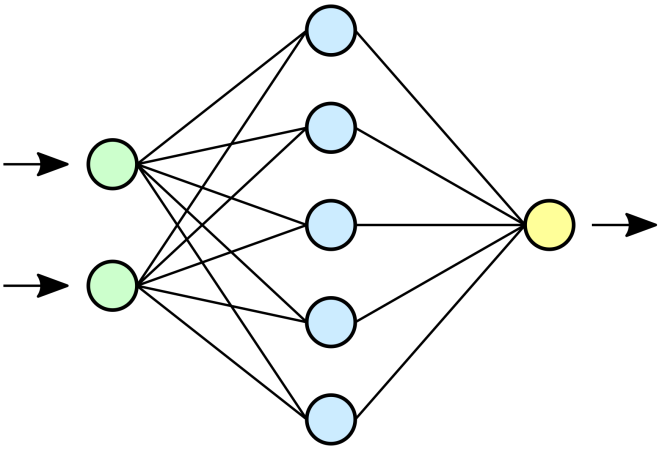


FIGURE 3.4 – Réseau de neurones simplifié et généralisé.

3.2.3.1 Perceptrons multicouches

En général, JO [78] est considéré comme le premier auteur à avoir utilisé un RN pour extraire les mots-clés d'un texte. Son approche est basée sur un perceptron multicouche (cachées). Chaque mot du texte à analyser (des articles scientifiques) est représenté par un vecteur de 3 valeurs réelles et de 3 valeurs booléennes (sous forme d'entiers égaux à 0 ou 1) : sa fréquence (TF), sa fréquence inverse de document (IDF), l'inverse de sa fréquence (ITF), s'il apparaît dans le titre du document, s'il apparaît dans la première phrase, s'il apparaît dans la dernière phrase. En sortie, le perceptron possède deux neurones qui permettent d'obtenir deux probabilités d'appartenir respectivement à la classe "mot-clé" ou "non mot-clé". Un vecteur de booléens à deux dimensions est construit à partir de ces valeurs et d'un seuil, où (1,0) signifie que le terme est un mot-clé et où (0,1) indique l'inverse. Les nombres de couches et de neurones cachés sont fixés arbitrairement par l'auteur. Les valeurs de précision obtenues montrent qu'avec une valeur de seuil proche de 0.5, cette méthode permet d'obtenir de meilleurs résultats que TF-IDF.

D'autres auteurs ont par la suite étendu les travaux de JO [78] mais en utilisant directement les probabilités comme scores pour classer les mots par ordre décroissant. Des caractéristiques différentes sont aussi utilisées dans les vecteurs en entrée :

— fréquence du terme (TF), fréquence inverse de document (IDF), rapport du nombre de paragraphes où le terme apparaît et du nombre de paragraphes du texte (PDF) et enfin s'il apparaît dans un titre (THS) pour WANG, PENG et HU [153] ;

— valeur TF-IDF du terme, position de la première occurrence normalisée et taille du mot ou de la phrase-clé pour Sarkar, Nasipuri et Ghose [136] ;

— mêmes que Jo [78] avec en plus la taille et la fréquence normalisées pour Aquino, Hasperué et Lanzarini [6].

Puisque les méthodes précédentes ne sont pas comparées entre elles, il est difficile de juger de l'impact positif ou négatif des différentes caractéristiques. Les valeurs de précision, rappel et f1-score obtenues par les différents auteurs s'accordent toutefois sur la supériorité de cette approche à base de perceptrons multicouches par rapport à la méthode supervisée KEA [158], basée sur une classification naïve bayésienne. Une étude comparative sur 210 articles scientifiques en anglais a été menée en 2012 par Sarkar, Nasipuri et Ghose [137] avec les mêmes représentations vectorielles des mots que dans leur méthode de 2010, citée précédemment [136]. Les valeurs de précision et de rappel obtenues montrent de nouveau une supériorité des perceptrons multicouches par rapport à la classification naïve bayésienne, mais aussi par rapport aux arbres de décision.

Une version plus évoluée des perceptrons multicouches, les *Deep Belief Networks* (DBN), est utilisée par Jo et Lee [79]. Ces RN apprennent automatiquement à reconstruire le vecteur qui leur est passé en entrée en identifiant les composants principaux du texte (les classes représentées par les neurones en sortie) et ce qui les caractérisent. Les DBN ont ainsi pour avantages d'être plus efficaces pour identifier les termes dont l'importance est moins explicite, comme les termes peu fréquents, et nécessitent de plus petits jeux d'entraînement que les perceptrons multicouches, pour des performances souvent équivalentes voire supérieures.

3.2.3.2 Réseaux de neurones récurrents

Les réseaux de neurones récurrents (*Recurrent Neural Networks* ou RNN) sont également étudiés pour extraire des mots-clés. Ils se caractérisent par l'utilisation de fonctions de combinaison non-linéaires, par la présence d'au moins un cycle dans la structure et par la non-nécessité d'utiliser un vecteur de taille fixe en entrée.

Un RNN profond avec couches jointes (*Joint-Layer Deep Recurrent Neural Networks* ou JL-DRNN) est employé par Zhang et al. [167], avec pour objectif de mieux détecter les mots-clés peu fréquents. En entrée, le réseau est alimenté par des représentations vectorielles à 300 dimensions des termes du texte, obtenues grâce à un modèle de plongement de mots pré-entraîné sur un jeu de données de Google. Deux couches cachées sont utilisées. La première permet d'identifier les mots-clés, la seconde le rôle que pourrait jouer le terme dans une phrase-clé (début, milieu, fin, non mot-clé, mot-clé seul). Ces deux couches cachées sont finalement jointes sur la couche de sortie. Les valeurs de précision, rappel et f1-score obtenues sur un jeu de données constitué de *tweets* (textes très courts) surpassent largement celles des méthodes non-supervisées, et dans une moindre mesure celles des méthodes supervisées.

Un RNN encodeur-décodeur est utilisé dans les méthodes de Meng et al. [107] et de Chen et al. [31], une nouvelle fois pour mieux détecter les mots-clés peu fréquents. Ces réseaux sont constitués de deux parties distinctes : l'encodeur, qui prend en entrée une séquence d'éléments de taille variable (ex : série temporelle, séquence de mots, etc.) et qui extrait pour chaque une représentation vectorielle de taille fixe ; et le décodeur, qui de manière traditionnelle utilise cette représentation pour classifier les mots. En laissant le système lui-même décider des caractéristiques importantes, les auteurs évitent de donner trop d'importance à juste une partie d'entre elles. La première méthode réussit à augmenter de manière significative le f1-score des principales

approches supervisées (KEA, Maui) et non-supervisées (TF-IDF, TextRank, SingleRank, ExpandRank), et dans certains cas même à le doubler. La seconde méthode ajoute des contraintes sur les segments-clés pour limiter la redondance et la mauvaise couverture des thématiques du texte, qui sont les principales faiblesses des résultats de la première méthode. De cette manière, tous les f1-scores obtenus sont cette fois-ci triplés. Toutefois, à cause de l'apprentissage non-supervisé des caractéristiques des mots par l'encodeur, ces RNN doivent traiter un très grand nombre de données avant de devenir efficaces.

Les réseaux de neurones récurrents bidirectionnels à mémoire court-terme et long terme (*Bidirectional Long Short-Term Memory* ou Bi-LSTM) sont exploités par Basaldella et al. [11] et par Alzaidy, Caragea et Giles [5]. Ils permettent de répondre au problème de disparition de gradient, qui empêche la mémorisation des évènements passés lors de l'apprentissage. Les méthodes proposées transforment l'extraction en un problème d'étiquetage de chaînes de caractères, comme le proposaient déjà certaines méthodes supervisées [165, 54]. La seconde méthode propose deux étiquettes différentes (mot-clé, non mot-clé), alors que la première trois (début de segment-clé, composant de segment-clé, non mot-clé). Les deux approches utilisent le modèle pré-entraîné de plongement de mots Glove [123] pour obtenir les représentations vectorielles des termes. Les valeurs de rappel obtenues dépassent largement celles des méthodes non-supervisées, et celles de f1-score surpassent aussi celles des méthodes supervisées basées sur un étiquetage.

Un dernier type de RNN, le réseau de neurones antagoniste génératif, est exploré par Wang et al. [155]. À l'origine à destination de la génération d'images, cette famille de réseaux est adaptée par les auteurs pour l'extraction automatique de mots-clés. Leur proposition, *Topic-based Adversarial Neural Network* (TANN), permet d'obtenir de meilleures valeurs de f1-score que toutes les méthodes basées RNN présentées dans cette partie. Toutefois, son bon fonctionnement requiert une nouvelle fois un très grand nombre d'extractions sur des textes différents.

3.2.3.3 Conclusion sur les réseaux de neurones

Pour conclure, les méthodes d'extraction de mots-clés basées sur les RN utilisent des **représentations vectorielles** des mots d'un texte, calculées par le système ou non, pour ensuite apprendre à leur attribuer une **probabilité** d'appartenir à un ensemble binaire ou non de **classes** de sortie. Ces approches obtiennent de **meilleurs scores de précision, rappel et f1-score** que les méthodes supervisées ou non-supervisées. Malgré l'importance du choix des caractéristiques qui composent les vecteurs d'entrée sur les résultats, peu d'études sur leurs impacts positifs ou négatifs sont menées. Cela mène un certain nombre d'auteurs à utiliser des RN particuliers, qui **apprennent eux-mêmes à extraire les représentations** avant la classification. Cependant, ces méthodes nécessitent pour la plupart un **très grand nombre** d'extractions et donc **de données** avant de devenir efficaces, ce qui peut être une limite en fonction des jeux à disposition. Les méthodes d'extraction basées sur des RN restent toutefois très prometteuses, notamment grâce à leur capacité à **mieux détecter les segments-clés peu fréquents**. Des systèmes capables d'apprendre eux-mêmes les représentations tout en ne nécessitant pas de trop gros jeux de données, comme les DBN, pourraient constituer une avancée majeure dans le domaine.

3.2.4 Conclusion de l'état de l'art

Pour conclure cet état de l'art, l'extraction automatique de mots-clés est une tâche qui vise à identifier et à sélectionner les termes d'un texte qui **décrivent le mieux son contenu**. Pour ce faire, les caractéristiques locales ou globales des termes, basées sur des concepts comme leur

fréquence, leur position, leur entropie, ou certaines informations externes lexicales, syntaxiques ou sémantiques, apportées par la langue ou des bases de connaissances, sont exploitées par des méthodes :

— **non-supervisées**, essentiellement créées à partir d'heuristiques et d'observations, qui réussissent à extraire des résultats pertinents mais peinent à identifier certaines catégories de segments-clés, comme par exemple ceux qui sont peu fréquents ;

— **supervisées**, qui construisent un modèle de discrimination par apprentissage pour capturer les corrélations cachées et/ou invisibles entre certaines caractéristiques choisies arbitrairement et ainsi couvrir une plus grande variété de segments-clés ;

— **basées réseaux de neurones**, qui permettent d'obtenir les meilleures valeurs de précision, rappel et f1-score, pour certaines en apprenant elles-mêmes les bonnes caractéristiques à utiliser, mais qui nécessitent de grands jeux de données pour correctement fonctionner.

La pertinence d'une méthode d'AKE est très dépendante de la taille des textes à analyser, ainsi que de leur consistance structurelle (s'ils possèdent un format ou une structure connue), du nombre de thématiques différentes qu'ils abordent et de la corrélation de ces sujets. Par conséquent, même si sur le papier les approches basées réseaux de neurones semblent les plus efficaces grâce à leur plus grande capacité à capturer les corrélations cachées entre les caractéristiques d'un texte, parfois même en les apprenant d'elles-mêmes, une étude comparative des principales méthodes de chaque approche est nécessaire pour pouvoir déterminer celle qui fournira les meilleurs résultats sur un jeu de données précis.

3.3 Expérimentations

Dans cette section, quelques expériences sur l'extraction automatique de mots-clés menées dans le cadre de mon stage sont présentées avec leurs résultats. Mon objectif est de déterminer les approches et méthodes qui permettent d'obtenir les résultats les plus pertinents pour annoter sémantiquement les médias référencés dans les *tweets* relatifs à la vaccination contre la Covid-19.

3.3.1 Construction du jeu de données

Le jeu de données étudié est toujours celui de l'étude de cas de la section 2.5. Pour rappel, celui-ci contient plus de 18 millions de *tweets* en français contenant au moins un mot-clé lié à la Covid-19 et aux vaccins et a été collecté grâce à l'architecture Hydre [51] entre le 1^{er} décembre 2020 et le 31 mars 2021 (120 jours). Les *tweets* peuvent prendre la forme de *tweets* originaux, de *retweets* (partages), de citations (partages avec commentaire supplémentaire) ou de réponses.

Puisque l'objectif de ces expérimentations est de déterminer l'approche d'extraction la plus adaptée pour annoter sémantiquement les médias référencés dans les *tweets* de ce jeu à l'aide de mots-clés, ma première étape a donc consisté à créer le jeu de données constitué de ces médias. Pour ce faire, j'ai commencé par retirer tous les *retweets* pour ne garder que les éléments qui apportent du nouveau contenu et qui ne sont pas juste un partage pur. Parmi ceux-ci, seuls ceux qui référencent un ou plusieurs médias à l'aide d'hyperliens sont conservés, soit environ 31.5%. Pour finir, la liste des hyperliens sans doublon est extraite (1 034 783 valeurs).

J'ai ensuite créé un script ETL (*Extract-Transform-Load*) pour télécharger les pages web référencées par les URL à ma disposition. J'ai choisi de le développer en Scala avec le *framework* Spark, un moteur distribué très efficace et très simple d'utilisation de traitement de larges

volumes de données, ce qui était particulièrement adapté à ma situation. Le script récupère la liste des hyperliens puis accède au contenu HTML des pages référencées (*scraping*) avant de l'enregistrer dans un fichier sur le disque dur de la machine. Les documents sont conservés tels quels pour ne pas créer de biais en altérant leur contenu. Cependant, ils ne sont pas tous enregistrés dans le même dossier. En effet, même si théoriquement il serait possible de le faire, les temps d'accès et de recherche dans un tel répertoire seraient assez élevés, ce qui pourrait être un souci ensuite pour enrichir la base de données après analyses et extractions. De plus, il vaut mieux privilégier une solution capable de gérer le passage à l'échelle.

J'ai donc décidé de m'inspirer des bases de données pour créer une arborescence adaptée à ma problématique. J'ai opté pour l'organisation des données dite aléatoire, qui propose des temps d'accès et d'écriture en O(1) (temps constant), et qui consiste à utiliser une fonction de hachage pour créer des adresses communes à des groupes d'éléments en fonction de leurs identifiants. Ce type de fonction prend en entrée une donnée de n'importe quelle taille et retourne en sortie une donnée de taille fixe (une séquence de bits, une chaîne de caractères, etc.). J'ai choisi la fonction *MurmurHash3*, qui produit des entiers de 32 bits bien distribués sur l'intervalle des valeurs possibles, pour ne pas me retrouver avec des répertoires avec 1 fichier et d'autres avec plusieurs dizaines de milliers. Chaque URL est hachée par la fonction et j'utilise le résultat pour construire mon arborescence : les 8 premiers bits fournissent l'identifiant du premier niveau, les 8 suivants celui du deuxième niveau, etc. Ce système me permet de répartir mes fichiers sur une arborescence à 4 niveaux de largeur maximale 255. Lorsqu'un accès au contenu d'un média est nécessaire, son URL a juste à être hachée par une fonction *MurmurHash3* puis son chemin dans l'arborescence peut être instantanément déduit.

En pratique, je n'ai pas téléchargé les 1 034 783 médias du jeu de données car il aurait été très coûteux en temps d'analyser une si grande quantité de données. Pour tout de même rester cohérent avec l'objectif que je m'étais fixé avec ces expériences, j'ai plutôt décidé de travailler sur un échantillon représentatif de la population. Ainsi, selon les méthodes d'échantillonage statistiques [3], seuls 385 fichiers devenaient nécessaires pour représenter tous les autres avec un niveau de confiance de 95% et une marge d'erreur de 5%, ce qui était déjà beaucoup plus abordable. J'ai donc lancé mon script ETL pour télécharger des fichiers choisis aléatoirement dans la liste complète des URL jusqu'à en obtenir le nombre suffisant. À chaque fois, un fichier n'était enregistré que s'il était exploitable par la suite, c'est-à-dire s'il respectait trois conditions :

- **être en français**, pour pouvoir utiliser des outils de TLN et pour être cohérent avec les *tweets* du jeu de base qui sont dans cette langue ;
- **posséder des mots-clés** assignés manuellement dans les métadonnées du fichier, pour pouvoir s'en servir ensuite comme mots-clés dorés lors des phases d'évaluation ;
- posséder au moins un mot-clé qui **apparaît dans le contenu** du texte (ceux qui ne le sont pas sont retirés).

Pour vérifier ces conditions, j'ai exploité la structure HTML et plus précisément les attributs *lang* et *content* des balises *<html>* et *<meta>* de type *keywords*.

Après création du jeu de données, j'ai découvert que l'assignation manuelle des mots-clés n'était pas toujours correctement effectuée (résumés entiers à la place des mots-clés, mots-clés vides, etc.). En tout, seuls 206 fichiers étaient annotés de la bonne manière. Après vérification, je me suis rendu compte que cette taille d'échantillon était toujours significative avec un niveau

3. `https://www.economie.gouv.fr/files/fiche_pratique_constitution_echantillonv1.pdf?fbclid=IwAR0i4MBS4gNWBoUHP6PfmtqkM3PVa7bN49vBHlivcpRNc8YJlh8KiGAjeKI`

de confiance de 95%, mais pour une marge 2% plus grande à 7% (taille > 196). J'ai donc décidé de simplement supprimer les fichiers exploitables mais malformés du jeu de données.

Pour finir, afin de pouvoir entraîner les méthodes supervisées, un jeu d'entraînement est nécessaire. La documentation de la bibliothèque Python *scikit-learn*, que j'ai ensuite utilisée pour mes expériences, recommande [4] de posséder un jeu d'entraînement 3 fois plus grand que le jeu de validation. J'ai donc relancé la même procédure que précédemment pour récupérer 618 fichiers supplémentaires, en veillant à ne pas considérer les médias déjà téléchargés.

Afin d'altérer au minimum la sémantique des documents, peu de pré-traitements sont appliqués sur les jeux. Le texte est passé en minuscules pour minimiser les variants morphologiques ("Covid-19", "covid-19"). Pour éviter les problèmes, les caractères spéciaux hors ponctuations sont retirés, notamment les retours à la ligne.

À la fin de cette première étape, je dispose donc de deux jeux de données, créés à partir des médias référencés dans les *tweets* du corpus portant sur la Covid-19 et la vaccination. Le premier, le jeu de validation, est constitué de 206 documents HTML et a pour but d'évaluer la précision de tous les types de méthodes, supervisées ou non. Le second, le jeu d'entraînement, est constitué de 618 documents HTML et est à destination des phases d'apprentissage des méthodes supervisées. En moyenne, les documents téléchargés sont composés de 800 mots. La taille médiane est de 539 mots.

3.3.2 Expériences

Mon but est de déterminer l'approche la plus adaptée au jeu de validation, et donc par extension au jeu de données de départ. Pour cela, je propose d'évaluer les performances (précision, rappel et f1-score) de 4 méthodes non-supervisées et de 4 méthodes supervisées (dont une approche réseaux de neurones).

Pour mener à bien ces expériences, j'ai choisi le langage Python pour sa simplicité et sa flexibilité d'utilisation ainsi que pour le grand nombre de bibliothèques disponibles, qui permettent de plus se concentrer sur les résultats que sur les méthodes en elles-mêmes. Pour toutes les méthodes qui nécessitent des outils de TLN, j'ai utilisé la bibliothèque *spacy* couplée avec le modèle "fr_core_news_sm" [5] qui est pré-entraîné sur des sites de *news* avec des textes de petites à moyennes tailles, ce qui correspond globalement au profil type des documents de mes jeux de données.

Pour les méthodes non-supervisées, j'ai choisi d'étudier les résultats de TF-IDF [81], YAKE ! [28], KP-Miner [14] et PositionRank [46]. Des implémentations de ces méthodes sont proposées dans la bibliothèque PKE [6] de Florian Boudin, co-auteur de TopicRank [24].

Pour les méthodes supervisées, j'ai fait le choix de ne pas utiliser certaines méthodes comme KEA ou GenEx mais plutôt de comparer directement les algorithmes d'apprentissage entre eux avec la même représentation vectorielle en entrée. Ainsi, je compare les résultats obtenus par classification naïve bayésienne, avec un arbre de décision renforcé par *bagging*, avec les SVM et avec un perceptron multicouche. Des implémentations de tous ces algorithmes sont proposées par la bibliothèque *scikit-learn* [7]. Pour les représentations vectorielles des mots, j'ai décidé de

4. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html?highlight=train_test_split#sklearn.model_selection.train_test_split
5. https://spacy.io/models/fr
6. https://github.com/boudinfl/pke
7. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

coupler les caractéristiques les plus mises en avant dans les différentes méthodes de l'état de l'art. Avant d'être passés aux méthodes d'apprentissage, les vecteurs sont normalisés par leur taille car certaines (notamment les perceptrons multicouches) sont très sensibles aux variations d'échelle entre les valeurs des différentes dimensions. Chaque terme candidat est représenté par les informations suivantes :

— son vecteur de plongement calculé par le modèle de *spacy*, constitué de 97 dimensions qui capturent à la fois sa sémantique et son rôle syntaxique ;

— sa fréquence ;

— sa valeur TF-IDF, qui mesure sa spécificité ;

— la position de sa première occurrence, normalisée par rapport à la taille du texte ;

— la position de sa dernière occurrence, normalisée de la même manière ;

— l'écart-type des distances de ses occurrences ;

— sa centralité de proximité.

Malgré la présence d'une thématique principale commune aux documents de mon corpus (voir section 3.2.1.1), j'utilise TF-IDF pour mieux détecter les sous-thématiques. La fréquence seule permet de compenser le manque de spécificité des termes des thèmes principaux. L'utilisation des informations de position est motivée par les hypothèses sur la répartition des segments-clés, notamment utilisées par Herrera et Pury [68] et par El-Beltagy et Rafea [14].

Pour évaluer la centralité de proximité des termes, je construis un graphe non-dirigé et non-pondéré de co-occurrences par phrase (voir section 3.2.1.3). Dans celui-ci, je ne conserve que les sommets qui correspondent à des noms, adjectifs ou nombres. Pour conserver un nombre constant de dimensions par mot, la valeur 0 est attribuée aux sommets retirés. Les natures grammaticales sont déterminées par *spacy*. Puisque j'utilise seulement la centralité de proximité, le poids des arêtes n'a pas d'importance. Le choix de cette centralité plutôt qu'une autre est motivé par l'envie de donner plus de caractéristiques utiles aux documents courts, puisque les autres sont plus efficaces sur les documents longs. Or, cette mesure est réputée comme la plus efficace sur ce type de documents [1, 22]. L'objectif derrière l'utilisation d'une caractéristique basée graphe est d'exploiter au mieux la structure sous-jacente du texte créée par les relations entre les termes.

Pour chaque méthode exécutée, je sélectionne les N candidats les plus importants. La valeur N correspond au nombre de mots-clés dorés associés au document. Pour les méthodes non-supervisées, l'importance est évaluée par la pondération attribuée par la méthode, par exemple la centralité des termes dans un graphe. Pour les méthodes supervisées, elle est évaluée par la probabilité d'appartenir à la classe des mots-clés.

Toutes les expériences (exécution des différentes méthodes dans les conditions énoncées dans cette sous-section) ont été réalisées sur un *notebook* Jupyter [8], une interface de programmation accessible par navigateur où des blocs de code peuvent être ajoutés, modifiés, supprimés et exécutés dynamiquement. Une mémoire commune entre les blocs est présente et les sorties des algorithmes restent affichées après exécution, ce qui permet de compartimenter et de mettre en forme facilement des flux de travaux (*workflows*) d'analyse. Jupyter propose plusieurs langages de programmation dont Scala, R et donc Python.

8. `https://jupyter.org/`

3.3.3 Méthode d'évaluation

Pour évaluer les résultats des méthodes d'extraction, j'ai décidé d'utiliser deux systèmes en parallèle : un premier à base de correspondances partielles (CP) et un second à base de correspondances exactes (CE). Dans le premier cas, un mot-clé est considéré comme vrai positif s'il est inclus ou s'il contient un des mots-clés dorés. Dans le second, un vrai positif est un mot-clé identique à un mot-clé doré au caractère près.

À cause de la méthodologie choisie pour déterminer le nombre de candidats à sélectionner, je ne peux pas évaluer le rappel (et donc le f1-score) des méthodes. En effet, la liste des mots-clés identifiés fait la même taille que la liste des mots-clés dorés pour chaque document. Par conséquent, les nombres de faux positifs et de faux négatifs sont toujours identiques (différence de la taille de la liste et du nombre de vrais positifs) et donc les valeurs de précision et de rappel aussi. Pour surpasser ce problème, il aurait fallu laisser chaque méthode sélectionner un nombre arbitraire de mots-clés. Dans le cas des méthodes supervisées, il aurait été possible de conserver tous les termes dont la probabilité dépassait un certain seuil. Pour les méthodes non-supervisées, une possibilité aurait été de calculer le score moyen par terme et de sélectionner ceux avec une valeur supérieure à la somme de la moyenne et de l'écart-type.

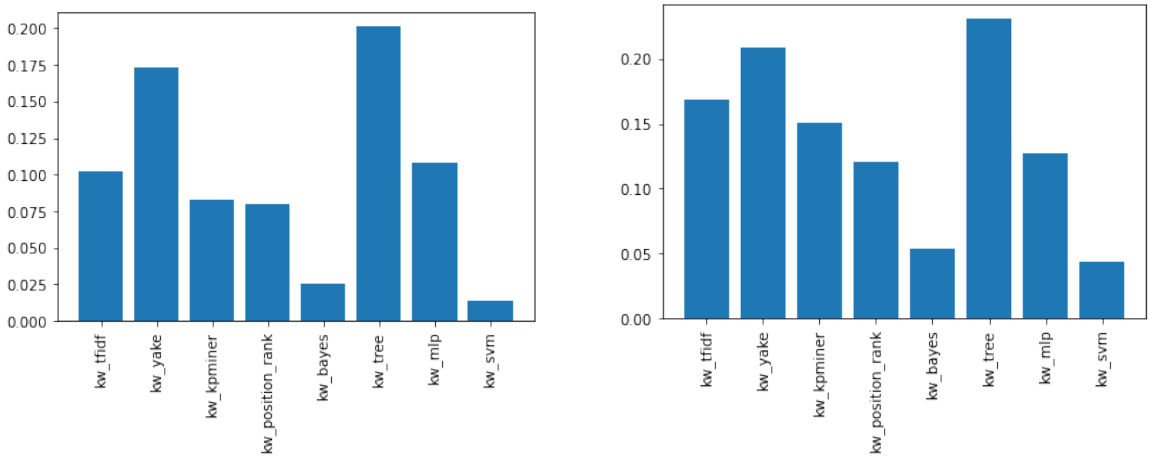
Pour des raisons de temps, je n'ai pas pu vérifier la robustesse et la fiabilité de tous mes modèles avec une validation croisée. Son principe est de découper le jeu de données en k blocs plutôt qu'en un jeu d'entraînement et un jeu de validation. L'entraînement est effectué k fois en sélectionnant à chaque fois un bloc pour évaluer les performances et les $k-1$ autres pour entraîner le modèle, sans modifier les paramètres de l'algorithme entre temps. La moyenne des k valeurs de précision obtenues permet d'estimer une valeur plus précise car moins sujette au risque de surapprentissage du modèle. J'ai pu vérifier la robustesse de ma classification naïve bayésienne, qui permet de correctement classifier 82% des mots en tant que mots-clés ou non.

3.3.4 Résultats

Les temps d'exécution et les valeurs de précision obtenues par les différentes méthodes sur le jeu de validation sont présentées dans la table 3.1. Quelle que soit la méthode d'évaluation choisie, les arbres de décision renforcés par *bagging* permettent d'obtenir les résultats les plus précis. Cependant, on remarque qu'ils sont suivis de près par la méthode non-supervisée YAKE!, sans jeu d'entraînement et pour un temps d'exécution deux fois moins important. Plusieurs raisons peuvent expliquer les résultats globalement moins bons des méthodes supervisées : taille du jeu, caractéristiques choisies pour la représentation vectorielle, mauvais paramétrage, etc. Le temps d'exécution très élevé et les résultats très mauvais de l'approche SVM sont dus au très grand nombre de dimensions des représentations vectorielles, ainsi qu'à la proportion beaucoup plus élevée de contre-exemples (mots qui ne sont pas des mots-clés), toutes les deux des faiblesses reconnues de cet algorithme d'apprentissage. Pour finir, on remarque que contrairement à la plupart des méthodes, les précisions CP des arbres de décision et du perceptron multicouche ne sont pas beaucoup plus élevées que leurs précisions CE, ce qui signifie que ces approches permettent de mieux cerner ce que la personne avait exactement en tête en attribuant les mots-clés dorés, sans approximation.

Méthode	**Précision CP**	**Précision CE**	**Temps d'exécution**
TF-IDF	16.9%	10.2%	32.4s
YAKE!	20.9%	17.3%	25m 38s
KP-Miner	15.1%	8.2%	26m 40s
PositionRank	12%	7.9%	26m 37s
Classif. naïve bayésienne	5.3%	2.5%	**10.3s**
Arbres décision + *Bagging*	**23.1%**	**20.1**%	42m 41s
SVM	4.4%	1.4%	1j 13h 45m
Perceptron multicouche	12.8%	10.8%	23m 9s

TABLE 3.1 – Performances des différentes méthodes sur le jeu de données.



(a) Correspondance exacte

(b) Correspondance partielle

FIGURE 3.5 – Performances des différentes méthodes sur le jeu de données.

3.3.5 Conclusion et limites

Pour conclure, les expérimentations menées dans cette section identifient les approches basées sur des arbres de décision renforcés par *bagging* comme méthodes d'extraction automatique de mots-clés les plus précises pour annoter sémantiquement les médias référencés dans les *tweets* du jeu de données du projet Cocktail. Toutefois, les très bons résultats obtenus par la méthode non-supervisée YAKE!, sans jeu d'entraînement et dans un temps d'exécution deux fois moins important, en font également un excellent candidat.

Ces résultats doivent être nuancés par quelques limites des expériences. D'abord, tous les mots-clés dorés attribués par les auteurs des différentes pages HTML ne sont pas de bonne qualité. La figure 3.6 montre quelques problèmes récurrents dans les listes de mots-clés dorés du jeu d'évaluation : redondance, recouvrement, utilisation de mots vides et/ou de verbes, etc. De plus, rien ne prouve que les termes choisis représentent bien le contenu du texte. Il faut garder à l'esprit qu'ils ont été avant tout choisis pour référencer une page web sur les moteurs de recherche. Ainsi, certains ont pu être mis en avant non pas à cause de leur importance mais

grâce à leur impact positif sur le référencement.

Ensuite, même avec une méthode d'évaluation par correspondance partielle, les résultats de précision sont très sensibles à la présence de variants, notamment morphologiques, mais aussi à l'utilisation de n-grammes. J'ai étudié la possibilité d'utiliser un outil de racinisation (*stemming*) pour réduire l'impact des variants, mais la mauvaise qualité de certains mots-clés dorés soulève de nombreuses questions : si les termes "vaccin" et "vaccination" sont tous les deux dorés, ont-ils une sémantique différente pour l'auteur du document ? Faut-il les considérer comme redondants à cause de leur racine commune ? Etc. J'ai donc choisi de ne pas utiliser de TLN dans mes évaluations. Tous ces éléments font que, dans les faits, une évaluation humaine des résultats obtiendrait sûrement des résultats différents.

Enfin, en guise de perspectives, d'autres expérimentations étaient à l'origine prévues pour cette partie. Une comparaison d'un plus grand nombre de réseaux de neurones différents aurait été intéressante, mais peu de bibliothèques Python proposent des implémentations de RN plus récents. Le module *keras* de *TensforFlow*[9] propose certains réseaux récurrents comme le Bi-LSTM, présenté dans la section 3.2.3.2. Cependant, la faible taille du jeu de données aurait très probablement été un problème avec ce type de RN. L'impact des caractéristiques choisies dans les représentations vectorielles des mots aurait également été un point intéressant à plus approfondir. Pour finir, la bibliothèque *spacy* propose des modèles pré-entraînés multilingues pour les tâches de TLN. Étudier leur efficacité pourrait permettre de s'affranchir de la barrière de la langue et ainsi pouvoir enrichir les *tweets* qui référencent des médias qui ne sont pas forcément en français.

	Gold
0	bush
1	vacciner
2	faire
3	caméra
4	même
5	disent
6	présidents
7	prêts
8	anciens
9	sont
10	obama
11	clinton
12	vaccin contre le coronavirus
13	devant
14	contre
15	les

	Gold
0	vaccin contre le covid-19
1	covid-19

FIGURE 3.6 – Exemples de mots-clés dorés de mauvaise qualité dans le jeu d'évaluation.

9. `https://keras.io/api/layers/recurrent_layers/`

3.4 Conclusion

Pour conclure sur l'annotation sémantique transmédia, mon stage m'a permis d'aborder la solution de l'extraction automatique de mots-clés pour enrichir les *tweets* référençant des documents majoritairement textuels. J'ai eu l'occasion d'étudier les travaux de recherche sur cette thématique au travers de la rédaction d'un état de l'art, qui présente plus de 90 méthodes réparties sur trois catégories : les approches non-supervisées, supervisées et basées sur des réseaux de neurones. Les approches non-supervisées pondèrent l'importance des termes en exploitant des caractéristiques qui leur sont inhérentes grâce à certaines formules statistiques ou grâce à des structures de données comme des graphes. Les approches supervisées identifient les mots-clés d'un texte grâce un algorithme d'apprentissage supervisé, par exemple un classifieur naïf bayésien, un arbre de décision ou un SVM. Enfin, les approches basées sur des réseaux de neurones, qui peuvent être supervisées ou non, utilisent ces systèmes pour mieux exploiter et détecter les corrélations cachées entre les caractéristiques des termes-clés. À l'aide des connaissances acquises par l'étude de la littérature, j'ai ensuite pu mettre en place quelques expériences pour départager les différentes approches sur un jeu de données du projet Cocktail.

Si les documents textuels représentent une plus grande partie des médias référencés dans les *tweets*, l'annotation sémantique transmédia concerne également le traitement d'images. L'affectation automatique de mots-clés (*Automatic Keyword Assignment* ou AKA) à partir d'un vocabulaire contrôlé type base de connaissance ou thésaurus est une solution à cette problématique et représente donc une perspective à mon travail de stage. Les méthodes d'apprentissage supervisées et les réseaux de neurones sont très utilisés en AKA grâce à leur capacité à classifier des éléments, les classes correspondant alors à des termes du vocabulaire. La bibliothèque *TensorFlow* (voir section 3.3) propose de nombreuses solutions intégrées pour traiter et classifier les images avec de telles approches.

Chapitre 4

Conclusion et Perspectives

En conclusion, j'ai eu l'occasion de travailler sur deux thématiques de recherche lors de mon stage : la détection de polarisation sur les réseaux sociaux et l'annotation sémantique transmédia. Sur le premier thème, j'ai rédigé deux articles scientifiques qui ont été acceptés par des conférences avec comité de lecture et un troisième est en cours de rédaction. J'ai également pu présenter oralement mes travaux lors de conférences en anglais et en français, lors de réunions du projet Cocktail et lors d'un séminaire du LIB. Sur les deux thématiques, j'ai pu à la fois proposer des contributions théoriques (formalisation de ma méthode, états de l'art, etc.) et pratiques (études de cas, expériences, etc.). Au milieu du stage, j'ai également pu participer à l'organisation de la conférence FRCCS 2021, ce qui m'a permis de découvrir un tout nouvel aspect du métier d'enseignant-chercheur.

D'un point de vue technique, le stage m'a permis de beaucoup travailler et de m'améliorer sur trois langages de programmation très populaires en sciences des données : R, Python et Scala. Mon travail sur la polarisation m'a permis d'approfondir mes connaissances sur la théorie et l'analyse des grands graphes et mon travail sur l'annotation sur les méthodes d'apprentissage supervisées ou non, notamment sur les différents types de réseaux de neurones, sur leur évaluation et sur la conception de jeux de données d'entraînement et de validation. J'ai également appris à me constituer moi-même un corpus d'articles scientifiques puis à m'en servir pour rédiger un état de l'art complet et mettre en place des expérimentations pour valider ou infirmer une théorie, compétences qui me seront très utiles lors de mon doctorat.

Par rapport à celui-ci, mon stage représente ainsi une parfaite introduction. D'un point de vue recherche, il m'a permis d'acquérir des connaissances, compétences et habitudes de travail qui me permettront d'aborder sereinement et efficacement les premières semaines de ma thèse. D'un point de vue thématique, l'extraction de métadonnées pour rendre les lacs de données sémantiques par l'utilisation de méthodes d'intelligence artificielle est une forme d'annotation sémantique. Elle passe par l'identification de caractéristiques et d'entités pour représenter les données semi ou non-structurées, ce qui correspond aux connaissances acquises. Ce stage aura donc été une excellente expérience professionnelle, capable à la fois de conclure un cycle et de servir de tremplin au suivant.

Bibliographie chapitre polarisation

[2] Md Tanvir Al Amin et al. « Unveiling polarization in social networks : A matrix factorization approach ». In : *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*. IEEE. 2017, p. 1-9.

[3] Andry Alamsyah et Fidocia Adityawarman. « Hybrid sentiment and network analysis of social opinion polarization ». In : *2017 5th International Conference on Information and Communication Technology (ICoIC7)*. IEEE. 2017, p. 1-6.

[8] Christopher A Bail et al. « Exposure to opposing views on social media can increase political polarization ». In : *Proceedings of the National Academy of Sciences* 115.37 (2018), p. 9216-9221.

[9] Albert-László Barabási et Eric Bonabeau. « Scale-free networks ». In : *Scientific american* 288.5 (2003), p. 60-69.

[18] Vincent D Blondel et al. « Fast unfolding of communities in large networks ». In : *Journal of statistical mechanics : theory and experiment* 2008.10 (2008), P10008.

[21] Stephen P Borgatti et Martin G Everett. « A graph-theoretic perspective on centrality ». In : *Social networks* 28.4 (2006), p. 466-484.

[25] Danah Boyd, Scott Golder et Gilad Lotan. « Tweet, tweet, retweet : Conversational aspects of retweeting on twitter ». In : *2010 43rd Hawaii international conference on system sciences*. IEEE. 2010, p. 1-10.

[34] Aaron Clauset. « Finding local community structure in networks ». In : *Physical review E* 72.2 (2005), p. 026132.

[35] Aaron Clauset, Mark EJ Newman et Cristopher Moore. « Finding community structure in very large networks ». In : *Physical review E* 70.6 (2004), p. 066111.

[37] Michael Conover et al. « Political polarization on twitter ». In : *Proceedings of the International AAAI Conference on Web and Social Media*. T. 5. 1. 2011.

[44] Michalis Faloutsos, Petros Faloutsos et Christos Faloutsos. « On power-law relationships of the internet topology ». In : *The Structure and Dynamics of Networks*. Princeton University Press, 2011, p. 195-206.

[49] Santo Fortunato. « Community detection in graphs ». In : *Physics reports* 486.3-5 (2010), p. 75-174.

[50] Kiran Garimella et al. « Quantifying controversy on social media ». In : *ACM Transactions on Social Computing* 1.1 (2018), p. 1-27.

[51] Annabelle Gillet, Éric Leclercq et Nadine Cullot. « Évolution et formalisation de la Lambda Architecture pour des analyses à hautes performances — Application aux données de Twitter ». In : *Revue ouverte d'ingénierie des systèmes d'information (ISI)* Numéro 1 (2021), p. 1-26.

[52] Michelle Girvan et Mark EJ Newman. « Community structure in social and biological networks ». In : *Proceedings of the national academy of sciences* 99.12 (2002), p. 7821-7826.

[53] Michel L Goldstein, Steven A Morris et Gary G Yen. « Problems with fitting to the power-law distribution ». In : *The European Physical Journal B-Condensed Matter and Complex Systems* 41.2 (2004), p. 255-258.

[55] Roberto González-Ibánez, Smaranda Muresan et Nina Wacholder. « Identifying sarcasm in Twitter : a closer look ». In : *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*. 2011, p. 581-586.

[56] Mark S Granovetter. « The strength of weak ties ». In : *American journal of sociology* 78.6 (1973), p. 1360-1380.

[58] Pedro Guerra et al. « A measure of polarization on social media networks based on community boundaries ». In : *Proceedings of the International AAAI Conference on Web and Social Media*. T. 7. 1. 2013.

[59] Pedro Guerra et al. « Antagonism also flows through retweets : The impact of out-of-context quotes in opinion polarization analysis ». In : *Proceedings of the International AAAI Conference on Web and Social Media*. T. 11. 1. 2017.

[61] Mohammad Nur Habibi et al. « Analysis of Indonesia Politics Polarization before 2019 President Election Using Sentiment Analysis and Social Network Analysis. » In : *International Journal of Modern Education & Computer Science* 11.11 (2019).

[73] Daniel J Isenberg. « Group polarization : A critical review and meta-analysis. » In : *Journal of personality and social psychology* 50.6 (1986), p. 1141.

[76] Julie Jiang, Xiang Ren et Emilio Ferrara. « Social media polarization and echo chambers : A case study of COVID-19 ». In : *arXiv preprint arXiv :2103.10979* (2021).

[82] Aditya Joshi, Pushpak Bhattacharyya et Mark J Carman. « Automatic sarcasm detection : A survey ». In : *ACM Computing Surveys (CSUR)* 50.5 (2017), p. 1-22.

[83] Andrea Kavanaugh et al. « Weak ties in networked communities ». In : *Communities and technologies*. Springer. 2003, p. 265-286.

[85] Jon M Kleinberg et al. « The web as a graph : Measurements, models, and methods ». In : *International Computing and Combinatorics Conference*. Springer. 1999, p. 1-17.

[89] Andrea Lancichinetti et Santo Fortunato. « Community detection algorithms : a comparative analysis ». In : *Physical review E* 80.5 (2009), p. 056117.

[91] Q Vera Liao et Wai-Tat Fu. « Beyond the filter bubble : interactive effects of perceived threat and topic involvement on selective exposure to information ». In : *Proceedings of the SIGCHI conference on human factors in computing systems*. 2013, p. 2359-2368.

[99] Shi-Long Luo, Kai Gong et Li Kang. « Identifying influential spreaders of epidemics on community networks ». In : *arXiv preprint arXiv :1601.07700* (2016).

[103] Matthew S McGlone. « Contextomy : the art of quoting out of context ». In : *Media, Culture & Society* 27.4 (2005), p. 511-522.

[109] Alfredo Jose Morales et al. « Measuring political polarization : Twitter shows the two sides of Venezuela ». In : *Chaos : An Interdisciplinary Journal of Nonlinear Science* 25.3 (2015), p. 033114.

[113] Mark EJ Newman. « Modularity and community structure in networks ». In : *Proceedings of the national academy of sciences* 103.23 (2006), p. 8577-8582.

[116] Tim O'reilly. *What is web 2.0.* " O'Reilly Media, Inc.", 2009.

[125] Pascal Pons et Matthieu Latapy. « Computing communities in large networks using random walks ». In : *International symposium on computer and information sciences*. Springer. 2005, p. 284-293.

[127] Usha Nandini Raghavan, Réka Albert et Soundar Kumara. « Near linear time algorithm to detect community structures in large-scale networks ». In : *Physical review E* 76.3 (2007), p. 036106.

[129] Jörg Reichardt et Stefan Bornholdt. « Statistical mechanics of community detection ». In : *Physical review E* 74.1 (2006), p. 016110.

[130] Howard Rheingold. *The Virtual Community, revised edition : Homesteading on the Electronic Frontier*. MIT press, 2000.

[133] Martin Rosvall, Daniel Axelsson et Carl T Bergstrom. « The map equation ». In : *The European Physical Journal Special Topics* 178.1 (2009), p. 13-23.

[143] Cass R Sunstein. « The Law of Group Polarization, 10 J ». In : *Pol. Phil* 175 (2002), p. 177.

[156] Douglas Brent West et al. *Introduction to graph theory*. T. 2. Prentice hall Upper Saddle River, 2001.

[157] Tamar Diana Wilson. « Weak ties, strong ties : Network principles in Mexican migration ». In : *Human Organization* (1998), p. 394-403.

[159] Zhihao Wu et al. « Efficient overlapping community detection in huge real-world networks ». In : *Physica A : Statistical Mechanics and its Applications* 391.7 (2012), p. 2475-2490.

[161] Jierui Xie, Stephen Kelley et Boleslaw K Szymanski. « Overlapping community detection in networks : The state-of-the-art and comparative study ». In : *Acm computing surveys (csur)* 45.4 (2013), p. 1-35.

Bibliographie chapitre annotation

[1] Willyan D Abilhoa et Leandro N De Castro. « A keyword extraction method from twitter messages represented as graphs ». In : *Applied Mathematics and Computation* 240 (2014), p. 308-325.

[4] Hassan Alrehamy et Coral Walker. « Exploiting extensible background knowledge for clustering-based automatic keyphrase extraction ». In : *Soft Computing* 22.21 (2018), p. 7041-7057.

[5] Rabah Alzaidy, Cornelia Caragea et C Lee Giles. « Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents ». In : *The world wide web conference*. 2019, p. 2551-2557.

[6] Germán Osvaldo Aquino, Waldo Hasperué et Laura Cristina Lanzarini. « Keyword extracting using auto-associative neural networks ». In : *XX Congreso Argentino de Ciencias de la Computación (Buenos Aires, 2014)*. 2014.

[7] Sanjeev Arora, Yingyu Liang et Tengyu Ma. « A simple but tough-to-beat baseline for sentence embeddings ». In : (2016).

[10] Regina Barzilay et Michael Elhadad. « Using lexical chains for text summarization ». In : *Advances in automatic text summarization* (1999), p. 111-121.

[11] Marco Basaldella et al. « Bidirectional lstm recurrent neural network for keyphrase extraction ». In : *Italian Research Conference on Digital Libraries*. Springer. 2018, p. 180-187.

[12] Slobodan Beliga. « Keyword extraction : a review of methods and approaches ». In : *University of Rijeka, Department of Informatics, Rijeka* (2014), p. 1-9.

[13] Slobodan Beliga, Ana Meštrović et Sanda Martinčić-Ipšić. « Toward selectivity based keyword extraction for Croatian news ». In : *arXiv preprint arXiv :1407.4723* (2014).

[14] Samhaa R El-Beltagy et Ahmed Rafea. « KP-Miner : A keyphrase extraction system for English and Arabic documents ». In : *Information systems* 34.1 (2009), p. 132-144.

[15] Kamil Bennani-Smires et al. « Simple unsupervised keyphrase extraction using sentence embeddings ». In : *arXiv preprint arXiv :1801.04470* (2018).

[16] Santosh Kumar Bharti et Korra Sathya Babu. « Automatic keyword extraction for text summarization : A survey ». In : *arXiv preprint arXiv :1704.03242* (2017).

[17] David M Blei, Andrew Y Ng et Michael I Jordan. « Latent dirichlet allocation ». In : *the Journal of machine Learning research* 3 (2003), p. 993-1022.

[19] Phillip Bonacich. « Power and centrality : A family of measures ». In : *American journal of sociology* 92.5 (1987), p. 1170-1182.

[20] Monali Bordoloi et al. « Keyword extraction using supervised cumulative TextRank ». In : *Multimedia Tools and Applications* 79.41 (2020), p. 31467-31496.

[22] Florian Boudin. « A comparison of centrality measures for graph-based keyphrase extraction ». In : *Proceedings of the sixth international joint conference on natural language processing*. 2013, p. 834-838.

[23] Adrien Bougouin. « Indexation automatique par termes-clés en domaines de spécialité ». Thèse de doct. Nantes, 2015.

[24] Adrien Bougouin, Florian Boudin et Béatrice Daille. « Topicrank : Graph-based topic ranking for keyphrase extraction ». In : *International joint conference on natural language processing (IJCNLP)*. 2013, p. 543-551.

[26] Sergey Brin et Lawrence Page. « The anatomy of a large-scale hypertextual web search engine ». In : *Computer networks and ISDN systems* 30.1-7 (1998), p. 107-117.

[27] Ronald S Burt. *Structural holes*. Harvard university press, 1992.

[28] Ricardo Campos et al. « Yake ! collection-independent automatic keyword extractor ». In : *European Conference on Information Retrieval*. Springer. 2018, p. 806-810.

[29] Jaime Carbonell et Jade Goldstein. « The use of MMR, diversity-based reranking for reordering documents and producing summaries ». In : *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, p. 335-336.

[30] Pedro Carpena et al. « Level statistics of words : Finding keywords in literary texts and symbolic sequences ». In : *Physical Review E* 79.3 (2009), p. 035102.

[31] Jun Chen et al. « Keyphrase generation with correlation constraints ». In : *arXiv preprint arXiv :1808.07185* (2018).

[32] Ping-I Chen et Shi-Jen Lin. « Automatic keyword prediction using Google similarity distance ». In : *Expert Systems with Applications* 37.3 (2010), p. 1928-1938.

[33] Gobinda G Chowdhury. « Natural language processing ». In : *Annual review of information science and technology* 37.1 (2003), p. 51-89.

[36] Jonathan D Cohen. « Highlights : Language-and domain-independent automatic indexing terms for abstracting ». In : *Journal of the American society for information science* 46.3 (1995), p. 162-174.

[38] John M Conroy et Dianne P O'leary. « Text summarization via hidden markov models ». In : *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 2001, p. 406-407.

[39] Sally F Dennis. « The design and testing of a fully automatic indexing-searching system for documents consisting of expository text ». In : *Information Retrieval : a Critical Review, Washington DC : Thompson Book Company* (1967), p. 67-94.

[40] Pedro Domingos et Michael Pazzani. « On the optimality of the simple Bayesian classifier under zero-one loss ». In : *Machine learning* 29.2 (1997), p. 103-130.

[41] Kathrin Eichler et Günter Neumann. « DFKI KeyWE : Ranking keyphrases extracted from scientific articles ». In : *Proceedings of the 5th international workshop on semantic evaluation*. 2010, p. 150-153.

[42] Gonenc Ercan et Ilyas Cicekli. « Using lexical chains for keyword extraction ». In : *Information Processing & Management* 43.6 (2007), p. 1705-1714.

[43] Günes Erkan et Dragomir R Radev. « Lexrank : Graph-based lexical centrality as salience in text summarization ». In : *Journal of artificial intelligence research* 22 (2004), p. 457-479.

[45] Jiajia Feng et al. « Keyword extraction based on sequential pattern mining ». In : *proceedings of the third international conference on internet multimedia computing and service*. 2011, p. 34-38.

[46] Corina Florescu et Cornelia Caragea. « Positionrank : An unsupervised approach to keyphrase extraction from scholarly documents ». In : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2017, p. 1105-1115.

[47] Corina Florescu et Wei Jin. « Learning feature representations for keyphrase extraction ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 32. 1. 2018.

[48] Edward W Forgy. « Cluster analysis of multivariate data : efficiency versus interpretability of classifications ». In : *biometrics* 21 (1965), p. 768-769.

[54] Sujatha Das Gollapalli, Xiao-Li Li et Peng Yang. « Incorporating expert knowledge into keyphrase extraction ». In : *Proceedings of the AAAI Conference on Artificial Intelligence*. T. 31. 1. 2017.

[57] Maria Grineva, Maxim Grinev et Dmitry Lizorkin. « Extracting key terms from noisy and multitheme documents ». In : *Proceedings of the 18th international conference on World wide web*. 2009, p. 661-670.

[60] Isabelle Guyon et al. « Gene selection for cancer classification using support vector machines ». In : *Machine learning* 46.1 (2002), p. 389-422.

[62] Yaakov HaCohen-Kerner. « Automatic extraction of keywords from abstracts ». In : *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer. 2003, p. 843-849.

[63] Yaakov HaCohen-Kerner, Zuriel Gross et Asaf Masa. « Automatic extraction and learning of keyphrases from scientific articles ». In : *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2005, p. 657-669.

[64] Mounia Haddoud et Saïd Abdeddaïm. « Accurate keyphrase extraction by discriminating overlapping phrases ». In : *Journal of Information Science* 40.4 (2014), p. 488-500.

[65] Kazi Saidul Hasan et Vincent Ng. « Automatic keyphrase extraction : A survey of the state of the art ». In : *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*. 2014, p. 1262-1273.

[66] Kazi Saidul Hasan et Vincent Ng. « Conundrums in unsupervised keyphrase extraction : making sense of the state-of-the-art ». In : *Coling 2010 : Posters*. 2010, p. 365-373.

[67] Marti Hearst. « What is text mining ». In : *SIMS, UC Berkeley* 5 (2003).

[68] Juan P Herrera et Pedro A Pury. « Statistical keyword detection in literary corpora ». In : *The European Physical Journal B* 63.1 (2008), p. 135-146.

[69] Bao Hong et Deng Zhen. « An extended keyword extraction method ». In : *Physics Procedia* 24 (2012), p. 1120-1127.

[70] Chong Huang et al. « Keyphrase extraction using semantic networks structure analysis ». In : *Sixth International Conference on Data Mining (ICDM'06)*. IEEE. 2006, p. 275-284.

[71] Anette Hulth. « Improved automatic keyword extraction given more linguistic knowledge ». In : *Proceedings of the 2003 conference on Empirical methods in natural language processing*. 2003, p. 216-223.

[72] JB Keith Humphreys. « Phraserate : An html keyphrase extractor ». In : *Dept. of Computer Science, University of California, Riverside, California, USA, Tech. Rep* (2002).

[74] Md Rafiqul Islam et Md Rakibul Islam. « An improved keyword extraction method using graph based random walk model ». In : *2008 11th International Conference on Computer and Information Technology*. IEEE. 2008, p. 225-229.

[75] Steven R Jefferts et al. « Accuracy evaluation of NIST-F1 ». In : *Metrologia* 39.4 (2002), p. 321.

[77] Xiao-yu Jiang. « Chinese automatic text summarization based on keyword extraction ». In : *2009 First International Workshop on Database Technology and Applications*. IEEE. 2009, p. 225-228.

[78] Taeho Jo. « Neural based approach to keyword extraction from documents ». In : *International conference on computational science and its applications*. Springer. 2003, p. 456-461.

[79] Taemin Jo et Jee-Hyong Lee. « Latent keyphrase extraction using deep belief networks ». In : *International Journal of Fuzzy Logic and Intelligent Systems* 15.3 (2015), p. 153-158.

[80] Thorsten Joachims. « Training linear SVMs in linear time ». In : *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006, p. 217-226.

[81] Karen Sparck Jones. « A statistical interpretation of term specificity and its application in retrieval ». In : *Journal of documentation* (1972).

[84] Su Nam Kim, Timothy Baldwin et Min-Yen Kan. « Evaluating n-gram based evaluation metrics for automatic keyphrase extraction ». In : *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*. 2010, p. 572-580.

[86] Mikalai Krapivin et al. « Keyphrases extraction from scientific documents : improving machine learning approaches with natural language processing ». In : *International Conference on Asian Digital Libraries*. Springer. 2010, p. 102-111.

[87] Bruce Krulwich et Chad Burkey. « Learning user information interests through extraction of semantically significant phrases ». In : *Proceedings of the AAAI spring symposium on machine learning in information access*. T. 25. 27. Stanford, CA. 1996, p. 110.

[88] Shibamouli Lahiri, Sagnik Ray Choudhury et Cornelia Caragea. « Keyword and keyphrase extraction using centrality measures on collocation networks ». In : *arXiv preprint arXiv :1401.6571* (2014).

[90] Alon Lavie et Michael J Denkowski. « The METEOR metric for automatic evaluation of machine translation ». In : *Machine translation* 23.2-3 (2009), p. 105-115.

[92] Don R Lick et Arthur T White. « k-Degenerate graphs ». In : *Canadian Journal of Mathematics* 22.5 (1970), p. 1082-1096.

[93] Chin-Yew Lin. « Rouge : A package for automatic evaluation of summaries ». In : *Text summarization branches out*. 2004, p. 74-81.

[94] Marina Litvak et al. « DegExt—A language-independent graph-based keyphrase extractor ». In : *Advances in intelligent web mastering–3*. Springer, 2011, p. 121-130.

[95] Zhiyuan Liu et al. « Automatic Keyphrase Extraction via Topic Decomposition ». In : *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA : Association for Computational Linguistics, oct. 2010, p. 366-376. URL : https://aclanthology.org/D10-1036.

[96] Zhiyuan Liu et al. « Clustering to find exemplar terms for keyphrase extraction ». In : *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, p. 257-266.

[97] Patrice Lopez et Laurent Romary. « HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID ». In : *SemEval 2010 Workshop*. 2010, 4-p.

[98] Hans Peter Luhn. « A statistical approach to mechanized encoding and searching of literary information ». In : *IBM Journal of research and development* 1.4 (1957), p. 309-317.

[100] Debanjan Mahata et al. « Theme-weighted ranking of keywords from text documents using phrase embeddings ». In : *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE. 2018, p. 184-189.

[101] Yutaka Matsuo et Mitsuru Ishizuka. « Keyword extraction from a single document using word co-occurrence statistical information ». In : *International Journal on Artificial Intelligence Tools* 13.01 (2004), p. 157-169.

[102] Yutaka Matsuo, Yukio Ohsawa et Mitsuru Ishizuka. « Keyworld : Extracting keywords from document s small world ». In : *International conference on discovery science*. Springer. 2001, p. 271-281.

[104] Olena Medelyan, Eibe Frank et Ian H Witten. « Human-competitive tagging using automatic keyphrase extraction ». In : *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009, p. 1318-1327.

[105] Olena Medelyan et Ian H Witten. « Thesaurus based automatic keyphrase indexing ». In : *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*. 2006, p. 296-297.

[106] S Menaka et N Radha. « Text classification using keyword extraction technique ». In : *International Journal of Advanced Research in Computer Science and Software Engineering* 3.12 (2013), p. 734-740.

[107] Rui Meng et al. « Deep keyphrase generation ». In : *arXiv preprint arXiv :1704.06879* (2017).

[108] Rada Mihalcea et Paul Tarau. « Textrank : Bringing order into text ». In : *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004, p. 404-411.

[110] Jane Morris et Graeme Hirst. « Lexical cohesion computed by thesaural relations as an indicator of the structure of text ». In : *Computational linguistics* 17.1 (1991), p. 21-48.

[111] Zara Nasar, Syed Waqar Jaffry et Muhammad Kamran Malik. « Named Entity Recognition and Relation Extraction : State-of-the-Art ». In : *ACM Computing Surveys (CSUR)* 54.1 (2021), p. 1-39.

[112] Zara Nasar, Syed Waqar Jaffry et Muhammad Kamran Malik. « Textual keyword extraction and summarization : State-of-the-art ». In : *Information Processing & Management* 56.6 (2019), p. 102088.

[114] Mark EJ Newman et Michelle Girvan. « Finding and evaluating community structure in networks ». In : *Physical review E* 69.2 (2004), p. 026113.

[115] Thuy Dung Nguyen et Min-Yen Kan. « Keyphrase extraction in scientific publications ». In : *International conference on Asian digital libraries*. Springer. 2007, p. 317-326.

[117] Yukio Ohsawa, Nels E Benson et Masahiko Yachida. « KeyGraph : Automatic indexing by co-occurrence graph based on building construction metaphor ». In : *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98-*. IEEE. 1998, p. 12-18.

[118] Miguel Ortuño et al. « Keyword detection in natural languages and DNA ». In : *EPL (Europhysics Letters)* 57.5 (2002), p. 759.

[119] Girish Keshav Palshikar. « Keyword extraction from a single document using centrality measures ». In : *International conference on pattern recognition and machine intelligence*. Springer. 2007, p. 503-510.

[120] Suhan Pan, Zhiqiang Li et Juan Dai. « An improved TextRank keywords extraction algorithm ». In : *Proceedings of the ACM Turing Celebration Conference-China*. 2019, p. 1-7.

[121] Kishore Papineni et al. « Bleu : a method for automatic evaluation of machine translation ». In : *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, p. 311-318.

[122] Claude Pasquier. « Single document keyphrase extraction using sentence clustering and latent dirichlet allocation ». In : *Proceedings of the 5th international workshop on semantic evaluation*. 2010, p. 154-157.

[123] Jeffrey Pennington, Richard Socher et Christopher D Manning. « Glove : Global vectors for word representation ». In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, p. 1532-1543.

[124] Matthew E Peters et al. « Deep contextualized word representations ». In : *arXiv preprint arXiv :1802.05365* (2018).

[126] J Ross Quinlan. *C4. 5 : programs for machine learning*. Elsevier, 2014.

[128] Iarivony Ramiandrisoa. « Extraction et fouille de données textuelles : application à la détection de la dépression, de l'anorexie et de l'agressivité dans les réseaux sociaux ». Thèse de doct. Université de Toulouse, Université Toulouse III-Paul Sabatier, 2020.

[131] Irina Rish et al. « An empirical study of the naive Bayes classifier ». In : *IJCAI 2001 workshop on empirical methods in artificial intelligence*. T. 3. 22. 2001, p. 41-46.

[132] Stuart Rose et al. « Automatic keyword extraction from individual documents ». In : *Text mining : applications and theory* 1 (2010), p. 1-20.

[134] Gerard Salton et Chris Buckley. « Automatic text structuring and retrieval-experiments in automatic encyclopedia searching ». In : *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*. 1991, p. 21-30.

[135] Gerard Salton, Anita Wong et Chung-Shu Yang. « A vector space model for automatic indexing ». In : *Communications of the ACM* 18.11 (1975), p. 613-620.

[136] Kamal Sarkar, Mita Nasipuri et Suranjan Ghose. « A new approach to keyphrase extraction using neural networks ». In : *arXiv preprint arXiv :1004.3274* (2010).

[137] Kamal Sarkar, Mita Nasipuri et Suranjan Ghose. « Machine learning based keyphrase extraction : comparing decision trees, naïve Bayes, and artificial neural networks ». In : *Journal of Information Processing Systems* 8.4 (2012), p. 693-712.

[138] Amit Singhal et al. « Modern information retrieval : A brief overview ». In : *IEEE Data Eng. Bull.* 24.4 (2001), p. 35-43.

[139] Min Song, Il-Yeol Song et Xiaohua Hu. « KPSpotter : a flexible information gain-based keyphrase extraction system ». In : *Proceedings of the 5th ACM international workshop on Web information and data management*. 2003, p. 50-53.

[140] Amy M Steier et Richard K Belew. « Exporting phrases : A statistical analysis of topical language ». In : *Second Symposium on Document Analysis and Information Retrieval*. Citeseer. 1993, p. 179-190.

[141] Lucas Sterckx et al. « Topical word importance for fast keyphrase extraction ». In : *Proceedings of the 24th International Conference on World Wide Web*. 2015, p. 121-122.

[142] Yi Sun et al. « SIFRank : a new baseline for unsupervised keyphrase extraction based on pre-trained language model ». In : *IEEE Access* 8 (2020), p. 10896-10906.

[144] Jie Tang et al. « Loss minimization based keyword distillation ». In : *Asia-Pacific Web Conference*. Springer. 2004, p. 572-577.

[145] Nedelina Teneva et Weiwei Cheng. « Salience rank : Efficient keyphrase extraction with topic modeling ». In : *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*. 2017, p. 530-535.

[146] Justine Raju Thomas, Santosh Kumar Bharti et Korra Sathya Babu. « Automatic keyword extraction for text summarization in e-newspapers ». In : *Proceedings of the international conference on informatics and analytics*. 2016, p. 1-8.

[147] Antoine Tixier, Fragkiskos Malliaros et Michalis Vazirgiannis. « A graph degeneracy-based approach to keyword extraction ». In : *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, p. 1860-1870.

[148] Peter D Turney. « Coherent keyphrase extraction via web mining ». In : *arXiv preprint cs/0308033* (2003).

[149] Peter D Turney. « Learning algorithms for keyphrase extraction ». In : *Information retrieval* 2.4 (2000), p. 303-336.

[150] Yasin Uzun. « Keyword extraction using naive bayes ». In : *Bilkent University, Department of Computer Science, Turkey www. cs. bilkent. edu. tr/˜ guvenir/courses/CS550/Workshop/Yasin_ pdf*. 2005.

[151] Didier A Vega-Oliveros et al. « A multi-centrality index for graph-based keyword extraction ». In : *Information Processing & Management* 56.6 (2019), p. 102063.

[152] Xiaojun Wan et Jianguo Xiao. « Single Document Keyphrase Extraction Using Neighborhood Knowledge. » In : *AAAI*. T. 8. 2008, p. 855-860.

[153] Jiabing Wang, Hong Peng et Jing-song Hu. « Automatic keyphrases extraction from document using neural network ». In : *Advances in Machine Learning and Cybernetics*. Springer, 2006, p. 633-641.

[154] Rui Wang, Wei Liu et Chris McDonald. « Using word embeddings to enhance keyword identification for scientific publications ». In : *Australasian Database Conference*. Springer. 2015, p. 257-268.

[155] Yanan Wang et al. « Exploiting topic-based adversarial neural network for cross-domain keyphrase extraction ». In : *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, p. 597-606.

[158] Ian H. Witten et al. « KEA : Practical Automatic Keyphrase Extraction ». In : *Proceedings of the Fourth ACM Conference on Digital Libraries*. DL '99. Berkeley, California, USA : Association for Computing Machinery, 1999, p. 254-255. isbn : 1581131453. doi : 10.1145/313238.313437. url : https://doi.org/10.1145/313238.313437.

[160] Fei Xie, Xindong Wu et Xingquan Zhu. « Efficient sequential pattern mining with wildcards for keyphrase extraction ». In : *Knowledge-Based Systems* 115 (2017), p. 27-39.

[162] Zhen Yang et al. « Keyword extraction by entropy difference between the intrinsic and extrinsic mode ». In : *Physica A : Statistical Mechanics and its Applications* 392.19 (2013), p. 4523-4531.

[163] Hongseon Yeom, Youngjoong Ko et Jungyun Seo. « Unsupervised-learning-based keyphrase extraction from a single document by the effective combination of the graph-based model and the modified C-value method ». In : *Computer Speech & Language* 58 (2019), p. 304-318.

[164] Wen-tau Yih, Joshua Goodman et Vitor R Carvalho. « Finding advertising keywords on web pages ». In : *Proceedings of the 15th international conference on World Wide Web*. 2006, p. 213-222.

[165] Chengzhi Zhang et al. « Automatic keyword extraction from documents using conditional random fields ». In : 4 (juin 2008), p. 1169-1180.

[166] Kuo Zhang et al. « Keyword extraction using support vector machine ». In : *international conference on web-age information management*. Springer. 2006, p. 85-96.

[167] Qi Zhang et al. « Keyphrase extraction using deep recurrent neural networks on twitter ». In : *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016, p. 836-845.

[168] Yongzheng Zhang, Evangelos Milios et Nur Zincir-Heywood. « A comparison of keyword-and keyterm-based methods for automatic web site summarization ». In : *AAAI04 Workshop on Adaptive Text Extraction and Mining*. 2004, p. 15-20.

Annexe A

Annexes

A.1 Étude de la corrélation entre l'antagonisme et la centralité HITS

Pour rappel, mon travail sur la polarisation m'a poussé à vouloir approfondir les possibles corrélations entre les indicateurs développés dans ma contribution et ceux traditionnellement utilisés en SNA, comme la modularité ou les différentes centralités. Au début du stage, plusieurs hypothèses ont été avancées et seule la première a pu être rapidement traitée : Les zones frontières d'une communauté contiennent majoritairement des hubs, les zones internes des autorités.

Cette hypothèse se base sur la centralité HITS. J'ai eu le temps de lancer quelques tests et d'étudier quelques résultats sans aboutir à de réelles conclusions intéressantes. La figure A.1 présente deux graphiques obtenus sur le jeu de données de l'étude de cas en appliquant l'algorithme HITS sur le graphe entier et en calculant les scores de hub et d'autorité moyens par zone. Dans les deux cas, les zones frontières semblent contenir des individus avec des centralités plus élevées que les zones internes. Ces résultats sont cependant difficiles à interpréter puisque le graphe est organisé en loi de puissance et que la moyenne entre tous les sommets n'a pas forcément beaucoup de sens.

Pour remédier à ce problème, les graphiques de la figure A.2 ont été générés. Cette fois-ci, une étiquette est attribuée à chaque sommet du graphe selon les règles suivantes : 'hub' si la valeur de hub est plus élevée que celle d'autorité, 'autorité' dans le cas inverse, aucune étiquette si égalité. J'ai ensuite calculé la proportion de hubs et d'autorités pour chaque zone (frontière ou interne) de chaque communauté. Une fois de plus, les zones frontières semblent centraliser à la fois les hubs et les autorités de chaque communauté. Cette vérité est cependant difficilement généralisable puisqu'elle varie en fonction des communautés. Il pourrait être intéressant de vérifier si d'autres éléments sont corrélés à ces variations (scores d'antagonisme, porosité, tailles des zones, etc.).

J'ai finalement décidé d'étudier le cas des utilisateurs les plus centraux selon HITS, qui sont répertoriés dans la figure A.3 en annexe. Les colonnes *boundary_with* et *internal_with* contiennent les identifiants des communautés avec lesquelles l'utilisateur est frontière ou membre interne. Pour ce cas particulier, les grandes figures d'autorité semblent effectivement plutôt faire partie des zones internes, comme le montrent l'absence valeurs dans la colonne *boundary_with* du tableau du dessus. Concernant les hubs, l'observation est un peu plus mitigée dans le sens où les utilisateurs les plus centraux semblent, en fonction des cas, se trouver soit en zone frontière

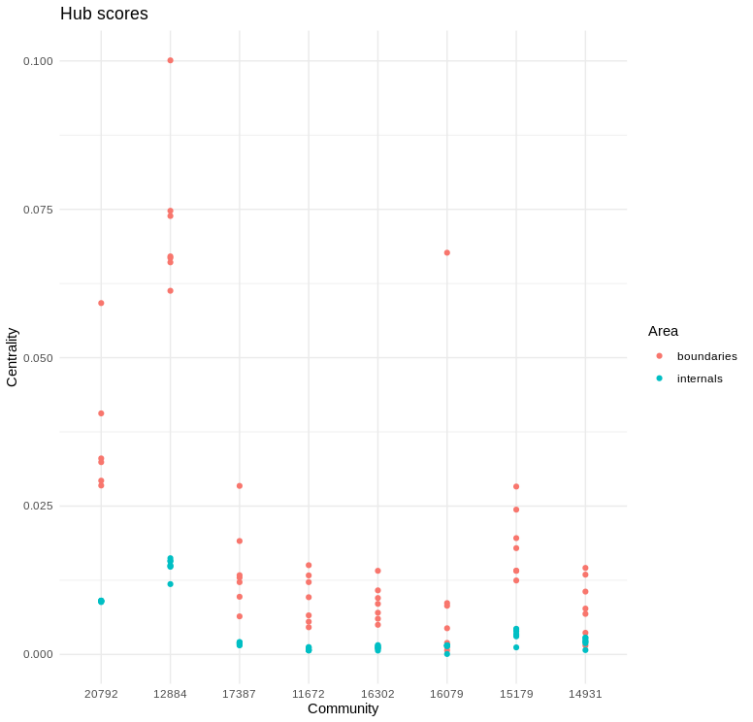
ou en zone interne, sans qu'une réelle tendance ne se dégage particulièrement.

En conclusion des quelques travaux effectuées pour vérifier l'hypothèse, les résultats des expériences ne permettent pas de confirmer ou d'infirmer l'hypothèse. Les autorités les plus centrales semblent effectivement se concentrer un peu plus dans les zones internes, mais les zones internes ne contiennent pas vraiment plus d'autorités que les zones frontières non plus. Pour ces dernières, il est vrai que les hubs principaux semblent se trouver souvent en frontière, sans que la tendance soit aussi nette que précédemment.

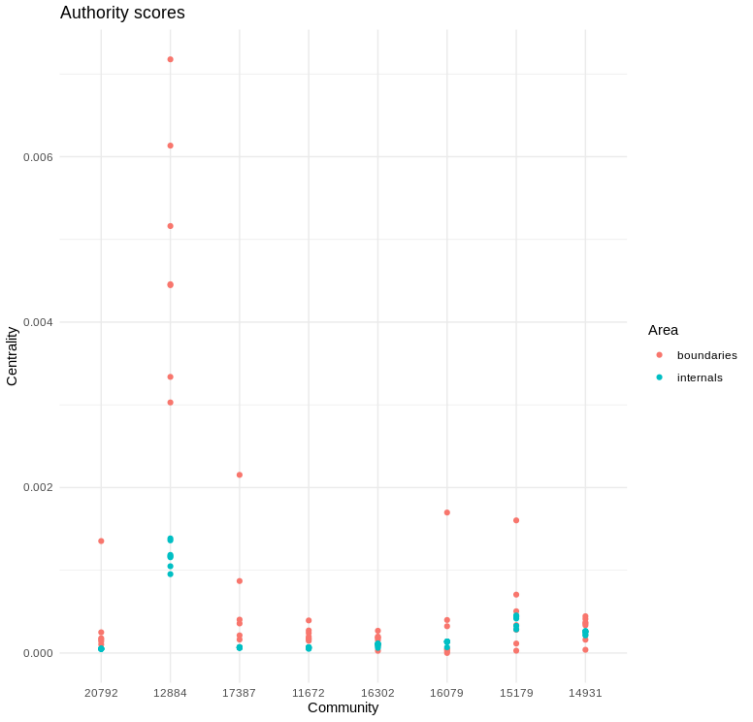
A.2 Tableaux et figures supplémentaires

Nom	**Taille**	**Type documents**	**Taille documents**	**Langue**
SemEval 2010		articles scientifiques	long	anglais
CogPrints		articles scientifiques	long	anglais
Sarkar, Nasipuri and Ghose (2012)	210	articles scientifiques	long	anglais
Turney (2000)	652	articles scientifiques, pages web, emails	long à court	anglais
ACM dataset	2 304	articles scientifiques	long	anglais
Wang, Peng and Hu (2006)	300	articles scientifiques	long	anglais, chinois
DEFT	93	articles scientifiques	long	français
DUC 2001	308	articles journalistiques	moyen	anglais
MPQA	535	articles journalistiques	moyen	anglais
Jo (2003)	900	articles journalistiques	moyen	anglais
20News	18 845	articles journalistiques	moyen	anglais
Sterckx, Demeester, Deleu and Develder		articles journalistiques	moyen	allemand
Inspec	2000	abstracts	court	anglais
OMD	3269	tweets	court	anglais

TABLE A.1 – Jeux de données les plus populaires en extraction automatique de mots-clés.

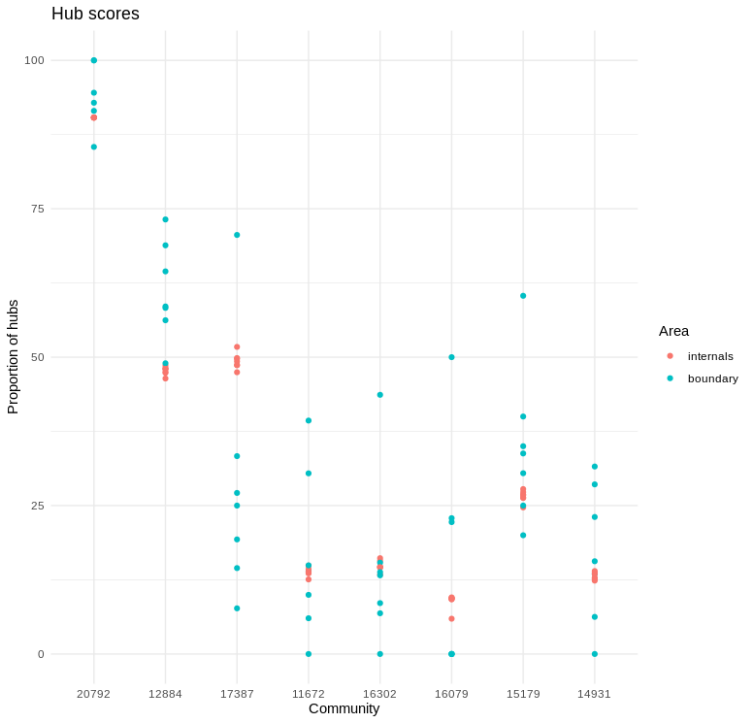


(a) score hub moyen

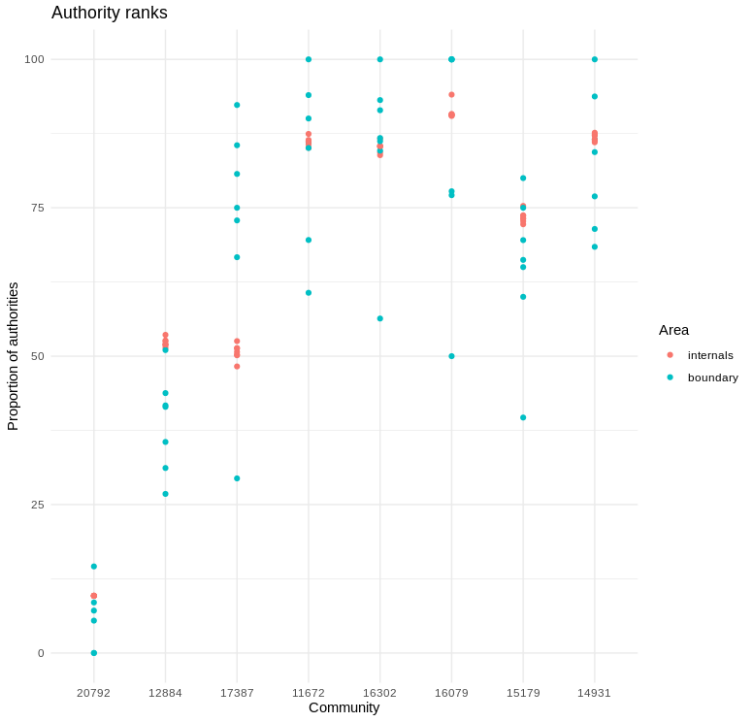


(b) score d'autorité moyen

FIGURE A.1 – Centralités moyennes des frontières et zones internes (cercles) et par communauté (colonnes).



(a) proportion de hubs



(b) proportion d'autorités

FIGURE A.2 – Proportions de hubs et d'autorités des frontières et zones internes (cercles) par communauté (colonnes).

Graphe des mentions – Top-K authorities

screen_name	name	verified	orientation	community	degree_in	degree_out	authority	hub	Pv	boundary_with	internal_with
silvano_trotta	SILVANO	FALSE	inconnu	12884	1548	40	1.00	0.10	0.47	15179	11672, 14931, 16079, 16302, 17387, 20792
Mediavenir	Mediavenir	NA	inconnu	20792	10271	1	0.48	0.02			11672, 12884, 14931, 15179, 16079, 16302, 17387
f_philippot	Florian Philippot	NA	inconnu	12884	871	3	0.28	0.01			11672, 14931, 15179, 16079, 16302
momotchiii	Momotchi	FALSE	inconnu	12884	617	7	0.27	0.21			11672, 14931, 15179, 16079, 16302, 17387, 20792
GaumontRene	LIBERTE	FALSE	inconnu	12884	571	3	0.18	0.16			11672, 14931, 15179, 16079, 16302, 17387, 20792
LeahCMa	Bibiel-Mahiout Leah	FALSE	inconnu	12884	62	75	0.18	0.56	0.46	11672, 14931, 15179, 16079, 16302, 17387	20792
DIVIZIO1	Fabrice Di Vizio	FALSE	antivax	12884	442	15	0.17	0.01	0.02	11672, 16302, 17387	14931, 15179, 16079, 20792
UPR_Asselineau	François Asselineau	TRUE	inconnu	12884	676	4	0.16	0.03	0.07	17387, 20792	11672, 14931, 15179, 16079, 16302
CaudeHenrion	Alexandra Henrion-Caude	FALSE	antivax	12884	442	1	0.14	0.00			11672, 14931, 15179, 16079, 16302, 20792
lolaweb71	lola cohen	NA	inconnu	12884	465	2	0.13	0.01			11672, 14931, 15179, 16079, 16302

Graphe des mentions – Top-K hub

screen_name	name	verified	orientation	community	degree_in	degree_out	authority	hub	Pv	boundary_with	internal_with
Tweet_Sopalinus	Tweet-Sopalinus	FALSE	inconnu	12884	1	154	0.00	1.00	0.48	11672, 15179, 16079, 16302, 17387, 20792	14931
Vally711	Vally 7	FALSE	provax	12884	0	32	0.00	0.95	0.49	15179, 16079	11672, 14931, 16302, 17387, 20792
DanielColloyald	Daniel Colloyald	FALSE	provax	12884	3	39	0.00	0.85	0.48	14931, 15179, 16079, 20792	11672, 16302, 17387
JPCodaccioni	Jean-Paul CODACCIONI	FALSE	inconnu	12884	0	48	0.00	0.79	0.47	15179, 16079, 17387	11672, 14931, 16302, 20792
FreemanFreestar	D. Francis D.	FALSE	inconnu	12884	0	2	0.00	0.72			11672, 14931, 15179, 16079, 16302, 17387, 20792
LeahCMa	Bibiel-Mahiout Leah	FALSE	inconnu	12884	62	75	0.18	0.56	0.46	11672, 14931, 15179, 16079, 16302, 17387	20792
delire67	délire67	FALSE	provax	12884	1	79	0.00	0.46	0.46	11672, 14931, 15179, 16079, 16302, 17387, 20792	
PRobesss	Pierre ROBESSS	FALSE	inconnu	12884	1	22	0.00	0.46	0.48	15179	11672, 14931, 16079, 16302, 17387, 20792
albert05759748	albert dupont	FALSE	inconnu	12884	3	101	0.00	0.44	0.47	11672, 14931, 15179, 16079, 16302, 17387, 20792	
RaderSerge	Serge Rader	FALSE	inconnu	12884	17	67	0.01	0.44	0.43	11672, 15179, 16302, 17387, 20792	14931, 16079

FIGURE A.3 – Utilisateurs avec les scores de hub et d'autorité les plus élevés du graphe.

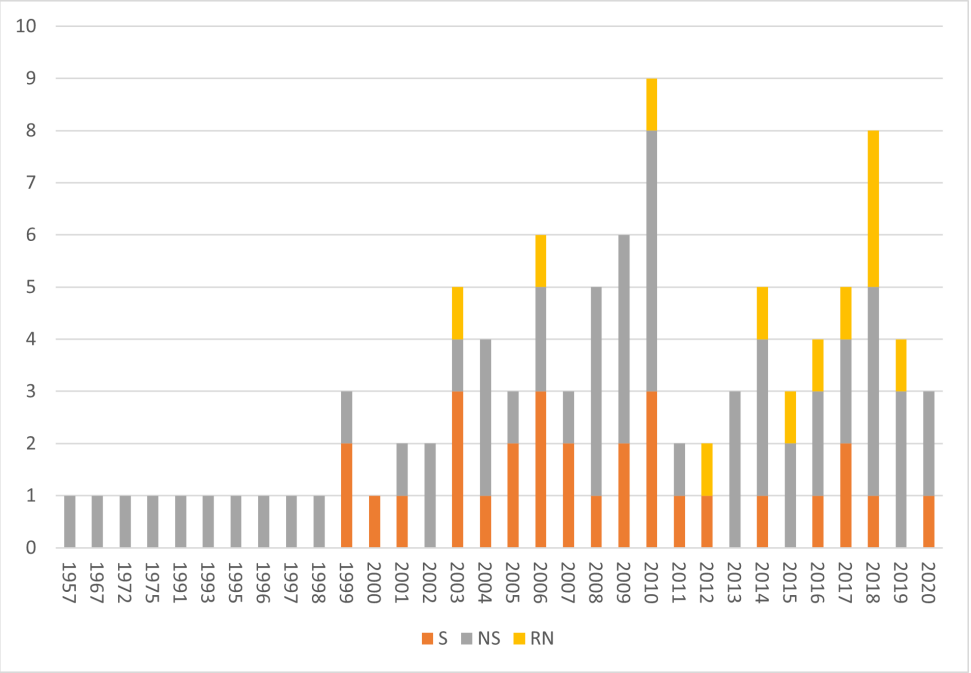


FIGURE A.4 – Nombre de méthodes d'extraction automatique de mots-clés par catégorie au cours du temps.

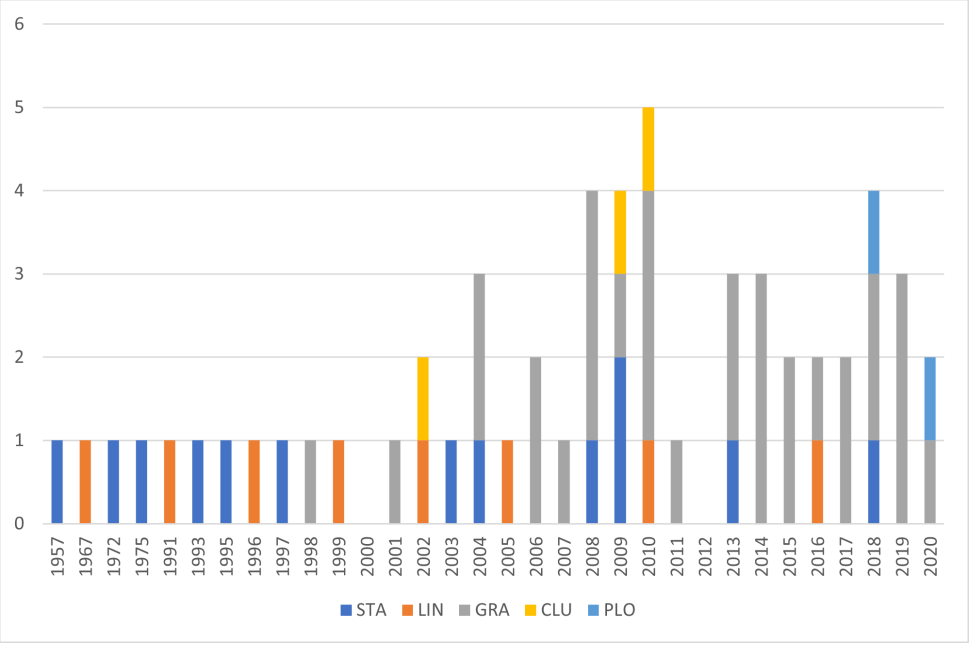


FIGURE A.5 – Nombre de méthodes d'extraction automatique de mots-clés non-supervisées par sous-catégorie au cours du temps.

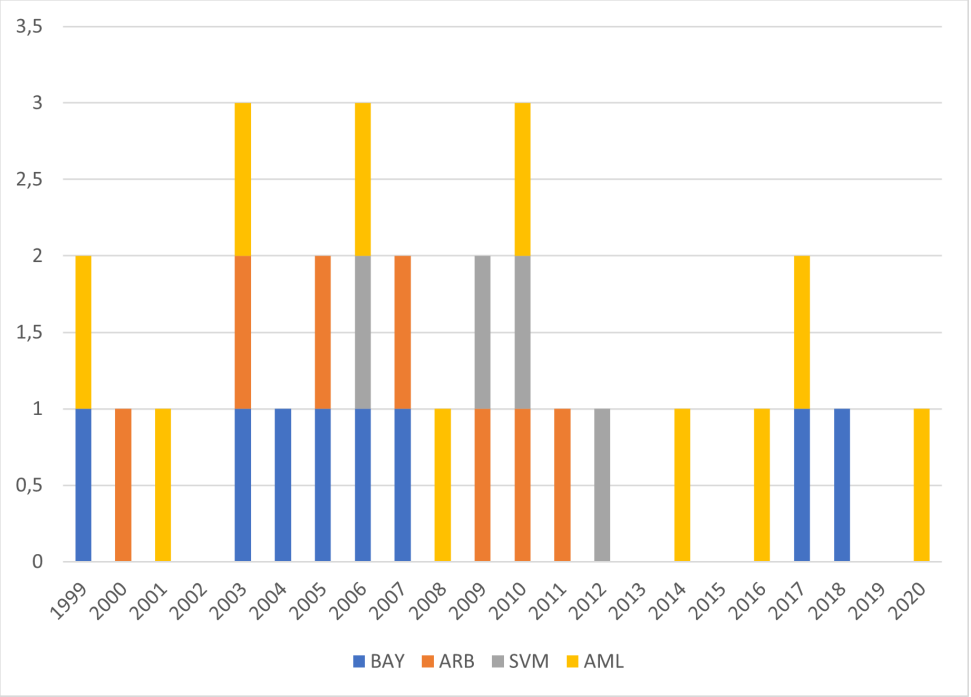


FIGURE A.6 – Nombre de méthodes d'extraction automatique de mots-clés supervisées par sous-catégorie au cours du temps.

Résumé

L'analyse des réseaux sociaux numériques (RSN) fait partie des problématiques actuelles majeures en sciences des données et a pour objectif l'étude de phénomènes, notamment sociologiques, grâce à l'extraction d'informations et de connaissances produites par les utilisateurs de ces applications au travers de la création d'un profil, des interconnexions avec d'autres personnes et de la publication de contenu.

De nombreux projets de recherche traitent aujourd'hui de cette thématique, comme par exemple le projet interdisciplinaire ISITE-BFC Cocktail, dans lequel s'inscrit mon stage et qui a pour ambition de mener à la création d'un observatoire en temps réel des tendances, des singularités et des signaux faibles circulant dans les discours de l'alimentaire et de la santé sur les RSN, et plus particulièrement sur Twitter, un réseau social dit de *microblogging* créé en 2006 et qui compte aujourd'hui plus de 199 millions d'utilisateurs quotidiens. La particularité principale de ce RSN réside dans le format des messages textuels pouvant être publiés sur l'application, les *tweets*, dont la taille est limitée à 280 caractères. Cette contrainte force une expression concise et souvent peu nuancée des avis des individus, facilitant l'étude hors contexte de l'environnement social présent sur le RSN.

La polarisation des utilisateurs est un exemple de phénomène pouvant intervenir sur les RSN et qui peut être identifié par l'étude des discours. Elle intervient lorsque les discussions autour d'un sujet particulier mènent des individus aux opinions similaires à se rassembler au sein de deux groupes principaux qui possèdent des positions opposées, conflictuelles et contrastées, alors que peu d'individus restent neutres ou dans une position intermédiaire. J'ai eu l'occasion de découvrir et de commencer à travailler sur cette thématique dans le cadre de mon projet tuteuré de master, et mes travaux sur le sujet ont mené à la rédaction d'un article en français accepté pour la conférence INFORSID'21. Les résultats obtenus et les retours très encourageants du comité de lecture de la conférence et des différents acteurs du projet Cocktail nous ont poussé avec Éric Leclercq à décider de consacrer une première partie du stage à prolonger le travail effectué jusque-là.

Dans un second temps, j'ai pu traiter une nouvelle problématique liée à l'utilisation des données de Twitter, l'enrichissement sémantique, qui consiste à exploiter au mieux les informations à notre disposition pour donner plus de sens aux messages très courts. L'objectif de cette partie du stage était alors d'étudier la possibilité d'annoter sémantiquement à l'aide de mots-clés les médias aux formats hétérogènes (images, vidéos, pages HTML, ...) référencés dans les *tweets*, notamment à partir de méthodes de *machine* ou de *deep learning*. Cette étude a mené à la rédaction d'un état de l'art et à l'élaboration d'un comparatif sur un jeu de données réelles collecté dans le cadre du projet Cocktail.