# Détection d'opinions, d'acteurs-clés et de communautés thématiques dans les médias sociaux

Guillaume Gadek

Normandie Université

# THÈSE

**Pour obtenir le diplôme de doctorat**

**Spécialité Informatique**

**Préparée au sein de l'INSA Rouen Normandie**

# Détection d'opinions, d'acteurs-clés et de communautés thématiques dans les médias sociaux

**Présentée et soutenue par**

**Guillaume GADEK**

| **Thèse soutenue publiquement le 22 novembre 2018 devant le jury composé de** | | |
|---|---|---|
| Cédric DU MOUZA | Maître de conférences HDR, *CEDRIC ; CNAM, Paris* | Rapporteur |
| Bénédicte LE GRAND | Professeur, *CRI ; Paris-1, Panthéon-Sorbonne* | Rapporteur |
| Florence SÈDES | Professeur, *IRIT ; Université P. Sabatier, Toulouse* | Examinateur |
| Rushed KANAWATI | Maître de conférences, *LIPN ; Paris-13, Viletaneuse* | Examinateur |
| Alexandre PAUCHET | Maître de conférences HDR, *LITIS ; INSA Rouen Normandie* | Directeur de thèse |
| Laurent VERCOUTER | Professeur, *LITIS ; INSA Rouen Normandie* | Examinateur, encadrant |
| Nicolas MALANDAIN | Maître de conférences, *LITIS ; INSA Rouen Normandie* | Invité, encadrant |
| Stéphan BRUNESSAUX | Senior Expert, *Airbus Defence and Space* | Invité, encadrant |
| Khaled KHELIF | Expert, *Airbus Defence and Space* | Invité, encadrant |

**Thèse dirigée par Alexandre PAUCHET et co-encadrée par Nicolas MALANDAIN et Laurent VERCOUTER**

AIRBUS INSA ROUEN NORMANDIE MI/IΣ litis

# Notice

# Remerciements

Issus d'un partenariat entre AIRBUS DEFENCE AND SPACE et le LITIS, ces travaux ont été réalisés dans le cadre d'une thèse CIFRE et bénéficie du soutien de l'ANRT : ce système, très pertinent, nécessite certes de trouver un équilibre ; en tout cas il m'a fait bénéficier du meilleur de deux mondes, l'académique et l'industriel.

Cette thèse a été rendue possible par la direction et l'accompagnement sans faille d'Alexandre Pauchet, et les conseils avisés de Laurent Vercouter et de Nicolas Malandain, mes co-encadrants. Malgré leurs responsabilités, ils ont su m'apprivoiser et me guider tout au long de ce parcours de recherche. Le côté industriel n'est pas en reste, et je tiens à remercier Stéphan Brunessaux, Khaled Khelif et Bruno Grilheres, qui ont su me pousser à saisir de nombreuses opportunités et m'ont aidé à surmonter les obstacles rencontrés.

Je tiens à remercier les membres du LITIS, et plus spécialement ceux de l'équipe MIND dont je fais partie, pour leur accueil chaleureux : les délicates invitations de Cecilia aux réunions d'équipe, les repas partagés avec Jean-Baptiste et Mathieu, les discussions en français, chinois et espagnol avec Qiushi et Franco, et surtout le soutien total de Sandra et Brigitte, notamment dans ma lutte contre les notes de frais.

L'équipe de traitement avancé de l'information, dirigée par Sylvie, m'a chaleureusement accueilli au sein d'AIRBUS, et je l'en remercie. Au cours de ces trois années, je l'ai vue croître, passer de quelques rescapés rolivalois à une équipe de choc, qui continue à croire en la recherche pour préparer le futur. Parmi cette équipe, je tiens à mentionner quelques piliers, prompts à refaire le monde autour d'une pinte (certains soirs) ou d'un café (tous les matins) : Jacques, Jonathan, Paul, Antoine, Kilian, Kévin, Kevin, et tous les autres qui font la vie de nos *open spaces*.

Enfin je ne peux manquer de mentionner ma famille, mes parents, mes neveux, et bien sûr mon amoureuse, Grecia ; un « merci » ne suffit pas à couvrir tout ce que je vis avec eux.

# Résumé

Les réseaux sociaux numériques ont pris une place prépondérante dans l'espace informationnel, et sont souvent utilisés pour la publicité, le suivi de réputation, la propagande et même la manipulation, que ce soit par des individus, des entreprises ou des états.

Alors que la quantité d'information rend difficile son exploitation par des humains, le besoin reste entier d'analyser un réseau social numérique : il faut dégager des tendances à partir des messages postés dont notamment les opinions échangées, qualifier les comportements des utilisateurs, et identifier les structures sociales émergentes.

Pour résoudre ce problème, nous proposons un système d'analyse en trois niveaux. Tout d'abord, l'analyse du message vise à en déterminer l'opinion. Ensuite, la caractérisation et l'évaluation des comptes utilisateurs est réalisée grâce à une étape de profilage comportemental et à l'étude de leur importance et de leur position dans des graphes sociaux, dans lesquels nous combinons les mesures topologiques d'importance des nœuds dans un graphe avec les statistiques d'engagement, par exemple en nombre d'abonnés. Enfin, le système procède à la détection et à l'évaluation de communautés d'utilisateurs, pour lesquelles nous introduisons des scores de cohésion thématique qui complètent les mesures topologiques classiques de qualité structurelle des communautés détectées.

Nous appliquons ce système d'analyse sur deux corpus provenant de deux médias sociaux différents : le premier est constitué de messages publiés sur Twitter, représentant toutes les activités réalisées par 5 000 comptes liés entre eux sur une longue période. Le second provient d'un réseau social basé sur TOR, nommé Galaxy2. Nous évaluons la pertinence de notre système sur ces deux jeux de données, montrant la complémentarité des outils de caractérisation des comptes utilisateurs (influence, comportement, rôle) et des communautés de comptes (force d'interaction, cohésion thématique), qui enrichissent l'exploitation du graphe social par les éléments issus des contenus textuels échangés.

# Abstract

Online Social Networks have taken a huge place in the informational space and are often used for advertising, e-reputation, propaganda, or even manipulation, either by individuals, companies or states.

The amount of information makes difficult the human exploitation, while the need for social network analysis remains unsatisfied: trends must be extracted from the posted messages, the user behaviours must be characterised, and the social structure must be identified.

To tackle this problem, we propose a system providing analysis tools on three levels. First, the message analysis aims to determine the opinions they bear. Then, the characterisation and evaluation of user accounts is performed thanks to the union of a behavioural profiling method, the study of node importance and position in social graphs and engagement and influence measures. Finally the step of user community detection and evaluation is accomplished. For this last challenge, we introduce thematic cohesion scores, completing the topological, graph-based measures for group quality.

This system is then applied on two corpora, extracted from two different online social media. The first is constituted of messages published on Twitter, gathering every activity performed by a set of 5,000 accounts on a long period. The second stems from a TOR-based social network, named Galaxy2, and includes every public action performed on the platform during its uptime. We evaluate the relevance of our system on these two datasets, showing the complementarity of user account characterisation tools (influence, behaviour and role), and user account communities (interaction strength, thematic cohesion), enriching the social graph exploitation with textual content elements.

# Table des matières

# Liste des figures

# Liste des tableaux

CHAPITRE 1

# Introduction

L'espace médiatique est une composante majeure de toute réussite ou échec, dans le domaine commercial ou de la défense. La réputation d'un produit vaut plus que ses caractéristiques techniques ; l'opinion publique peut transformer en défaite politique une suite de succès militaires sur le terrain. Dans cet espace, des adversaires prennent position, et conduisent des opérations d'influence et parfois même de désinformation. Or le paysage informationnel a été profondément modifié par le Web 2.0 et les réseaux sociaux, apparus au début des années 2000. La diffusion de l'information, auparavant limitée à quelques émetteurs (*broadcast* ou *one-to-many*), parvenant à atteindre toute la population via la presse écrite, la radio et la télévision, a évolué et inclut désormais de nombreux émetteurs (*multicast* ou *many-to-many*), aux influences variables.

Cet espace complexe est propice à l'obtention d'informations, de par sa diffusion très large dans la population, et de par son instantanéité. Il est aussi propice à des actions de manipulation et de désinformation de la part d'adversaires, tant dans le domaine économique que pour la défense, qui sont les deux grands domaines applicatifs considérés dans nos travaux. En conséquence, des outils sont nécessaires pour traiter de grandes quantités de documents collectés, et pour identifier les actions, souvent innovantes, de manipulation.

## 1.1 Contexte : les réseaux sociaux numériques

Afin de construire le contexte autour de ces travaux de thèse, cette section revient sur l'histoire des réseaux sociaux numériques, leurs spécificités et leur empreinte géographique. Leur importance économique est soulignée, ainsi que leur intérêt dans le domaine de la sécurité et de la défense.

### 1.1.1 L'essor du Web social

Le Web social ou 2.0 repose sur la possibilité donnée au lecteur, non propriétaire d'un site Web, d'y laisser un texte, commentaire, ou autre contenu. Il regroupe ainsi les commentaires d'articles de presse, les forums, et surtout les réseaux sociaux numériques, dont le premier exemple à succès est MySpace. Apparu en 2003, il permet à un utilisateur d'avoir sa propre page, et de créer des liens avec d'autres utilisateurs « amis ». MySpace fut ensuite déclassé par Facebook (créé en 2004) et sa fulgurante ascension. D'autres réseaux fonctionnent de la même façon mais se dédient à des domaines plus restreints, tels LinkedIn et Viadeo pour la vie professionnelle, ou encore ResearchGate pour l'activité académique.

Le Web social inclut aussi les forums, les blogs et le micro-blogging. Alors que les blogs sont personnels et ne contiennent des publications que d'un seul auteur, les sites de micro-blogging agrègent

les contenus produits par tous les utilisateurs. Les principaux exemples de micro-blogging incluent Instagram, Pinterest et surtout Twitter, fondé en 2006 et rendu célèbre par son côté public (aucune restriction d'accès aux messages, hors correspondance privée), son instantanéité, et sa contrainte des 140 caractères maximum par message. Twitter rassemble, mi-2018, 335 millions d'utilisateurs actifs. Si certains l'utilisent de manière quotidienne pour interagir avec leurs amis, d'autres s'en servent à des fins professionnelles : leur compte est la preuve de leur intérêt continu pour une thématique, une technologie, une problématique. Enfin, en France comme aux États-Unis, le rôle politique de Twitter s'est imposé : la plupart des élus y ont un compte, et s'en servent pour placer leurs éléments de discours et approcher les journalistes.

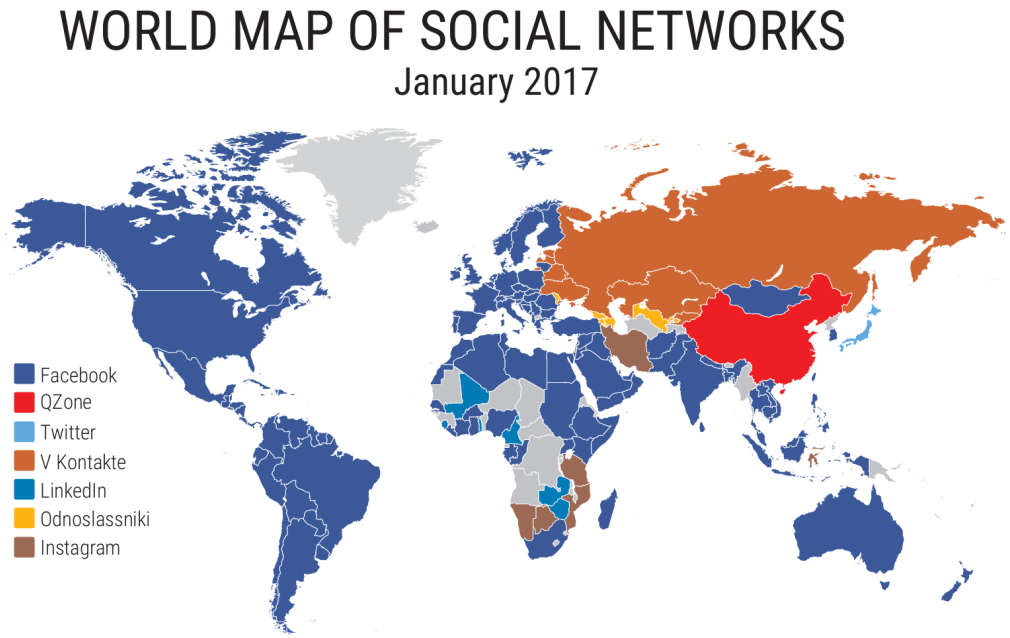



FIGURE 1.1 – Carte des réseaux sociaux les plus utilisés par pays en janvier 2017, `vincos.it/`

Il existe une géographie des réseaux sociaux, où l'extra-territorialité du Web n'est pas constatée partout. Facebook et Twitter dominent le paysage en Occident et au Moyen-Orient, avec quelques exceptions en voie de disparition : l'espagnol Tuenti, le brésilien Orkut ont, un temps, rivalisé avant de disparaître, comme le montrent les figures 1.1 et 1.2 qui proviennent des travaux de Vincenzo Cosenza [1]. En figure 1.1, la domination de Facebook se remarque ; seules la Russie et la Chine semblent y échapper, chacun protégeant son champion domestique. La figure 1.2 montre le second réseau le plus populaire, par pays. Ici, l'information manque pour de nombreux territoires. Le marché est plus partagé, avec l'Amérique latine qui se tourne vers Instagram, Twitter qui apparaît sur quelques pays, et les forums de Reddit qui émergent en Australie et au Canada.

La Russie et les pays de la Communauté des États Indépendants (ex-URSS) connaissent en effet un écosystème indépendant, reposant sur quelques géants russes : Yandex est plébiscité comme moteur de recherche, et Vkontakte [2], fondé en 2006 par Pavel Durov (désormais à la tête de Telegram), propose les fonctionnalités à la fois de Facebook et Youtube. Odnoklassniki [3] fonctionne sur un principe similaire, proposant des espaces de discussion, le partage de fichiers et média (photos, musiques et vidéos). Enfin, la Grande Muraille Numérique a complètement soumis le marché chinois aux acteurs intérieurs, Sina Weibo (équivalent de Twitter) et QQ (similaire à Facebook). L'intégration de ces géants dans le paysage réel est d'ailleurs impressionnante, puisque les consommateurs physiques paient la plupart de leurs transactions via leurs comptes de réseaux sociaux.

1. `vincos.it/`
2. Accessible sur `vk.com`
3. Accessible sur `ok.ru`

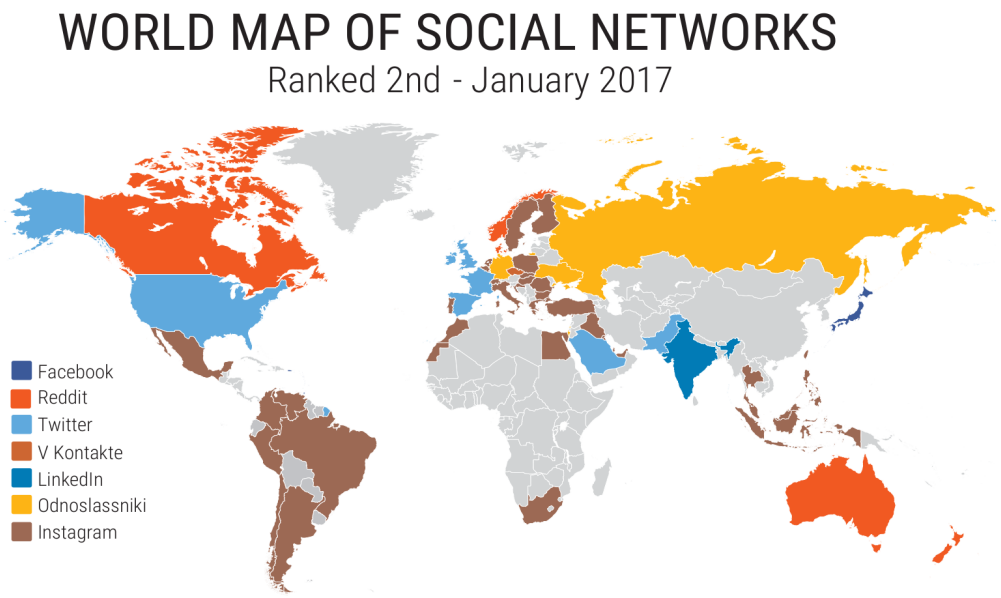



FIGURE 1.2 – Carte des seconds réseaux sociaux les plus utilisés par pays en janvier 2017, `vincos.it/`

L'analyse des réseaux sociaux répond ainsi à un véritable besoin de comprendre et d'identifier les phénomènes à l'œuvre sur l'espace informationnel. Une première application, à but lucratif, accompagne les départements marketing des entreprises souhaitant maintenir leur bonne réputation, et tisser des liens avec les internautes. Une seconde application est orientée vers la défense et la sécurité : il s'agit de combattre la cyber-criminalité, contrer la propagande ou les manipulations.

### 1.1.2 Intérêt économique

Les réseaux sociaux sont activement exploités par les entreprises, via leurs départements marketing. En effet, les plate-formes telles que Twitter ou Instagram permettent aux fabricants de dialoguer directement avec leurs clients, et ainsi d'identifier les usages de leurs produits en réalisant un retour d'expérience continuel. Ils permettent aussi d'identifier l'apparition de rumeurs, parfois néfastes. En miroir, ils sont exploités pour créer et diffuser des contenus promotionnels, et attirer les clients.

L'intelligence économique consiste à recueillir des informations, de manière légale et éthique, à propos des partenaires réels ou potentiels, des concurrents, et de l'environnement en général. Il faut identifier les risques autour de l'activité actuelle ou future de l'entreprise, détecter les nouvelles réglementations et lois impactant le fonctionnement de la compagnie. Cette activité repose sur l'identification de sources pertinentes d'information, leur lecture, et finalement la production de renseignement à fournir à l'équipe dirigeante. La presse, les ministères et territoires, les sites Web des concurrents sont des candidats pertinents, ainsi que leurs comptes respectifs, que ce soit sur Twitter, Instagram ou ailleurs. Face à la quantité de textes à lire et à maintenir à jour, le besoin de s'équiper se fait ressentir.

L'entreprise doit surveiller ses concurrents, mais aussi elle-même. L'apparition de rumeurs néfastes signifie parfois la fin de l'activité commerciale. À titre d'exemple, la sortie d'un jeu vidéo (StarWars Battlefront2), en novembre 2017, s'est accompagnée d'un fort mouvement de protestation sur le web social à l'encontre de l'éditeur *Electronics Arts*, ou EA : tout un pan du jeu nécessitait de payer des frais supplémentaires pour être accessible. Les contenus tournant l'éditeur en ridicule se sont accompagnés d'appels au boycott et d'insultes. EA n'a pas eu d'autre solution que de céder[4]. Il ne fait guère de doute que la virulence de la protestation et sa dimension sociale ont eu un impact,

4. `https://www.forbes.com/sites/erikkain/2017/11/28/ea-shares-plummet-after-star-wars-battlefront-ii-loot-box-fiasco/`

alors que les articles des médias spécialisés classiques ne semblaient pas scandalisés outre mesure par le modèle commercial.

Les réseaux sociaux sont maintenant pris en compte par les annonceurs pour la réalisation de campagnes publicitaires. La publication de contenus payants, similaires aux contenus gratuits habituels sur les plate-formes sociales, représente une grande part des budgets marketing : en 2016 en France, il s'agit d'un marché de presque un milliard d'euros. Le simple affichage payant n'est pas nouveau, et est assimilable à la location d'espaces publicitaires sur un site web ; il bénéficie toutefois d'une plus-value intéressante : le ciblage exploite les données des profils sociaux, dont le lieu de résidence, l'âge et la profession. Cependant, d'autres techniques sont aussi utilisées, avec notamment le recours à des influenceurs : des tiers qui vont, moyennant finances, mettre en valeur les produits de l'entreprise. La production du contenu publicitaire n'a plus à être déléguée à une agence de création de pubs ; l'influenceur s'en charge, satisfait d'avoir obtenu gracieusement les produits commentés ensuite par vidéo Youtube, article de blog, tweet ou encore post sur Instagram.

### 1.1.3 Intérêt dans le domaine de la défense et de la sécurité : manipulation

Ces techniques de persuasion commerciale sont parfois détournées à d'autres fins, plus politiques ou idéologiques. Militer pour un parti politique, même en ligne, est tout à fait éthique et légitime ; recourir à la manipulation et au mensonge pour diffuser la propagande d'une organisation terroriste ne l'est pas. Proclamé en 2014, le « califat » (Daech ou état islamique) était déjà doté d'une division numérique, considérant l'espace informationnel comme une province, un territoire, au même titre que l'Irak ou la Syrie. À ce titre ont été nommés en 2014-2015, des responsables de la province numérique, chargés de coordonner l'action et de gérer ce « territoire », avec des objectifs quantifiés : recrutement, diffusion géographique, résilience. De nombreuses études [5] se sont penchées sur l'action des terroristes en ligne : comment opèrent-ils, qu'y font-ils ? Où sont situés les comptes actifs ? Sur Twitter en 2016, on estimait à 100 000 le nombre de comptes utilisateurs, soutiens de Daech [NATO, 2016].

Les organisations terroristes n'ont pas l'apanage de l'action opérationnelle sur les réseaux sociaux. Des soupçons très lourds pèsent sur l'implication des services russes dans le déroulement de la campagne présidentielle américaine de 2016, qui a résulté en l'élection de Donald Trump. Au piratage des bases de données du serveur de mails du parti démocrate, divulguant le mode de fonctionnement interne du parti en pleine campagne [6], s'ajoute l'emploi massif de faux comptes sur les réseaux sociaux Facebook, Twitter et Reddit. Un autre type d'action suspecte passe par la diffusion de publicités sur les média sociaux, et surtout sur Facebook via des *mèmes* : souvent des images contenant du texte, modifiées et diffusées en masse [7]. Une courte liste de telles publicités est rendue publique par le parti démocrate [8] et suggère que les annonces ont été payées par des « entreprises proches du Kremlin ». Dans un second temps seulement sont apparues des traces moins conventionnelles, telle cette liste de comptes Twitter, bannis, mais estampillés comme de faux comptes manipulés depuis la Russie [9]. Cette attribution reste démentie par le Kremlin, globalement invérifiable. L'éventuelle existence d'une agence de désinformation russe [10] semble participer à la stratégie du pouvoir : il s'agit de montrer une capacité, qu'ils n'ont pas forcément. De même, les principales analyses sont menées par l'OTAN, qui a tout intérêt à grossir toute menace potentielle pour pouvoir l'anticiper et y répondre à temps, même quand la menace n'existe pas encore.

5. Voir notamment les publications du Centre contre le terrorisme : `https://www.ctc.usma.edu/about`
6. Le piratage est souvent attribué à la Russie : `https://www.washingtonpost.com/world/national-security/russian-government-hackers-penetrated-dnc-stole-opposition-research-on-trump/2016/06/14/cf006cb4-316e-11e6-8ff7-7b6c1998b7a0_story.html`
7. `https://fr.wikipedia.org/wiki/Mème_Internet`
8. `https://democrats-intelligence.house.gov/hpsci-11-1/`
9. `https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf`
10. `https://www.theguardian.com/media/2017/nov/20/russian-troll-army-tweets-cited-more-than-80-times-in-uk-media`

Cambridge Analytica revendique, sans preuve, avoir fait basculer le vote vers « *Leave* » lors du référendum pour le Brexit en juin 2016. Parmi leurs outils figurent la diffusion massive de contenus sur les réseaux sociaux, afin de contenir la « narrative » adverse et d'imposer la leur auprès des audiences. D'autres pays ont recours à de telles techniques, parfois de manière continue. Le quotidien britannique *The Guardian* exposait ainsi, en novembre 2016, qu'une trentaine d'états avaient recours à des « *opinions shapers* », formeurs d'opinions en ligne, pour occuper l'espace médiatique et légitimer le discours et l'action de leurs gouvernements [11]. Ces *opinions shapers* peuvent être déployés avec un coût et une complexité faibles : il suffit de rémunérer quelques centimes par message, par exemple via une plate-forme comme Amazon Mechanical Turk, en recourant à du *crowdsourcing*. Il n'y a alors aucun besoin de créer de nouveaux comptes ni de les insérer dans le réseau social global.

L'emploi massif d'étudiants, par exemple en Russie par la « *Internet Research Agency* » [12], a permis le pourrissage organisé des blogs et sites de presse, selon un ouvrage traitant de la désinformation en ligne [Giorgio Bertolin, 2017], produit par l'OTAN. Tout article est systématiquement commenté par des messages agressifs (« trolls ») envers les opposants ou modérés. Le résultat attendu de la politique de *trolling*, c'est que l'internaute, dégoûté de l'ambiance délétère sur le Net, n'aille plus poster de commentaire désapprobateur, que ce soit en ligne ou hors ligne. Si les citoyens n'osent plus faire de politique, il n'y a pas d'inquiétude à avoir de perdre une élection. Cette méthode déclenche un second phénomène, l'*astro-turfing*, qui consiste à affirmer la présence écrasante de son camp politique sur les réseaux sociaux, à revendiquer l'adhésion populaire sur le monde numérique. Cet objectif peut être atteint par l'utilisation de comptes à durée de vie très éphémère, à la popularité préparée à l'avance à l'aide d'une politique multi-comptes, avec des envois planifiés de messages [Shaheen, 2015].

Le domaine de la détection de prête-noms, ou faux comptes, manipulés sous une fausse identité, leur attribue le qualificatif de *sockpuppets*. Parfois aussi nommés *sybils*, ils atteignent des niveaux de raffinement très inégaux : certains comptes sont assez creux, créés à la va-vite sans photo de profil. D'autres paraissent tout à fait réalistes, disposent de liens sociaux crédibles, d'une cohérence géographique soignée. En 2012, Facebook estimait qu'il y avait au moins 83 millions de comptes *sybils* sur leur service [13]. Leurs usages sont divers : prise de position non assumée, donner une impression de nombre, vente de followers, collecte dissimulée de données ou encore attaques de *spam* et *phishing*. Le *botnet* découle des sockpuppets : c'est un réseau de robots. Plutôt que d'avoir une unique fausse identité, il s'agit d'un ensemble plus ou moins coordonné de faux profils, multipliant ainsi les chances d'atteindre un objectif fixé.

Les médias sociaux, tout comme le reste de notre environnement, sont propices aux actions et innovations de la part d'adversaires, dans le cadre de la loi ou au-delà. La masse d'information et la vitesse de publication rendent nécessaires des outils automatiques pour obtenir les informations nécessaires à temps et pour pouvoir reconnaître les tentatives de manipulations lorsqu'elles ont lieu.

## 1.2 Objectif : comprendre l'espace médiatique

Avec l'émergence des médias de masse au XX$^{eme}$ siècle, une réflexion fut menée sur leurs rôle et impact. En 1948, [Lasswell, 1948] posait la question « *who says what to whom in what channel with what effect* » : qui dit quoi, à qui, via quel canal, avec quel impact. Cette question reste toujours d'actualité, applicable à l'ensemble des canaux d'information dont nous disposons.

Concernant les médias sociaux et réseaux sociaux, nous pouvons enrichir la problématique par deux constats provenant d'approches bien distinctes. Une première observation vient du métier du renseignement d'origine sources ouvertes : quelle que soit l'information, elle est présente sur Internet. Il

11. https://www.theguardian.com/technology/2017/nov/14/social-media-influence-election-countries-armies-of-opinion-shapers-manipulate-democracy-fake-news

12. une agence équivalente existe en Chine, avec la « *Internet Water Army* » : https://en.wikipedia.org/wiki/Internet_Water_Army

13. https://www.forbes.com/sites/davidthier/2012/08/02/83-million-estimated-facebook-profiles-are-fake/

faut alors adapter les capteurs, collecteurs et processeurs pour tirer parti des données récupérables sur les réseaux sociaux, qui ne sont pas des sites Web comme les autres. De nombreuses problématiques apparaissent naturellement : comment récupérer les messages postés, comment identifier l'information parmi le bruit, comment classifier la langue d'un message, son thème, comment faire le lien entre plusieurs messages.

La seconde constatation, c'est qu'un réseau social numérique fonctionne comme un réseau social réel, comme notre société : il y a de nombreux acteurs, poursuivant des objectifs parfois proches, parfois opposés. Le réseau commente les événements du monde réel, mais il est aussi le lieu d'occurrence d'événements virtuels : des comptes naissent, vivent, se relient, se détachent, produisent, consomment, disparaissent. Les questions portent alors sur des éléments spécifiques au réseau social investigué : qui sont les créateurs d'un hashtag, qui sont les utilisateurs les plus en vue sur une thématique, comment fonctionnent les groupes de soutien à une marque, quand sont apparus les protestataires sur une nouvelle campagne de publicité.

Forts de ces constats, nous avons identifié trois axes de recherche que nous détaillons par la suite, qui constituent notre problématique : *Détection des opinions, des acteurs-clés, et des communautés thématiques dans les médias sociaux.*

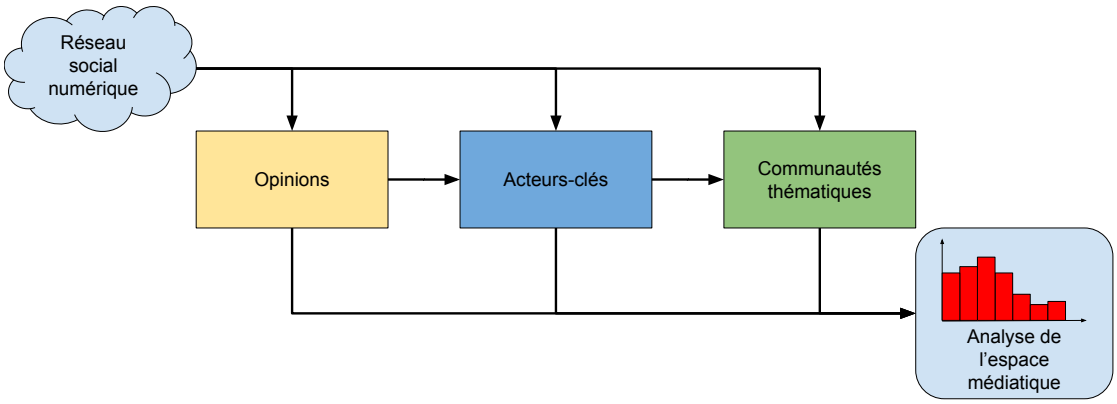



FIGURE 1.3 – Trois axes de recherche dans l'analyse des réseaux sociaux numériques

Notre approche est illustrée en figure 1.3, où un réseau social numérique constitue la donnée brute, ensuite analysée successivement au regard des opinions, des acteurs-clés et des communautés, chaque axe de recherche enrichissant le suivant. Le résultat final rassemble toutes les clés de lecture pour l'étude d'une plate-forme ou média social.

### 1.2.1 Détecter les opinions présentes dans les messages

Le premier axe de notre thèse porte sur l'analyse des textes présents sur les réseaux sociaux. Malgré la mode de l'image voire de la vidéo, qui permet à tout un chacun de créer et diffuser des photos, de les retoucher en quelques secondes, ou d'émettre une vidéo en direct, l'omniprésence des messages textuels demeure et soulève plusieurs défis dans ce domaine. L'un des défis majeurs consiste en la compréhension du texte, visant à rendre possible l'agrégation des opinions, entités nommées et vues exprimées sur l'ensemble des publications.

En effet, l'opinion permet de résumer le texte de manière satisfaisante : en marketing, ce sur quoi porte l'opinion permet d'identifier les préoccupations des consommateurs, alors que le sentiment ou polarité de l'opinion évalue l'adhésion ou le rejet du marché vis-à-vis d'un nouveau produit. En défense et sécurité, ce sur quoi porte l'opinion permet d'identifier les valeurs et peurs de la population. De son côté, la polarité de l'opinion évalue par exemple le soutien aux forces armées dans le cadre d'une opération extérieure. Des notions connexes, telles que l'émotion des locuteurs ainsi que la posture adoptée lors d'un débat présentent un intérêt analytique complétant les opinions détectées.

Il est nécessaire que le processus soit complètement automatisé, car la quantité de données est

considérable : sur Twitter par exemple, 500 millions de messages sont publiés chaque jour [14]. De plus, il s'agit d'une tâche répétitive et peu gratifiante. Nous avons ainsi décidé de couvrir les aspects de **détection de l'opinion** éventuellement contenue dans des messages issus de réseaux sociaux.

### 1.2.2 Identifier et caractériser les acteurs-clés

Le second axe de notre thèse porte sur l'analyse des profils utilisateurs, membres d'un réseau social. En effet, tout type de compte existe : certains remplissent consciencieusement tous les champs possibles, tels que les centres d'intérêt, ville, études, date de naissance, photos de profil, de bannière, de vacances... D'autres restent minimalistes et ne renseignent qu'un pseudonyme. Certains comptes sont robotisés, d'autres sont partagés et gérés par plusieurs personnes. Enfin, une entité (personne, entreprise, ...) peut se créer plusieurs comptes : personnel et professionnel, par segment de marché (par exemple un compte par pays ou par région).

Pour mettre en situation un message, le lecteur doit bénéficier d'informations concernant l'auteur du message. Il est peu probable qu'il connaisse l'ensemble des 300 millions d'utilisateurs actifs sur Twitter, ou encore les deux milliards de membres de Facebook. Un double travail est à mener ici, avec une première question portant sur l'estimation de l'impact d'un message, lié à l'audience et à l'influence de son auteur. Nous allons ainsi comparer plusieurs scores d'influence, qui permettent de se faire une idée rapide de la puissance de diffusion émanant d'un compte.

La seconde question, plus large, vise à caractériser l'auteur du message en exploitant toutes les informations disponibles. Cette caractérisation se décompose en plusieurs aspects, portant tant sur la localisation du compte, sur ses liens avec ses amis, sur ses données biographiques, que sur son comportement. Des indicateurs sont nécessaires pour estimer l'activité, le recours aux hashtags, photos, liens, quantifier la vie sociale le nombre de liens vers d'autres comptes, ou encore prendre en compte l'orthographe soignée (ou non) des messages. Nous plaçons ces éléments d'analyse comportementale, ainsi que les scores d'influence, sous l'intitulé de **détection d'acteurs-clés**.

### 1.2.3 Identifier et caractériser les communautés thématiques

Enfin, le troisième axe de notre thèse porte sur l'un des aspects les plus passionnants des réseaux sociaux : l'émergence de structures sociales, d'ensemble d'utilisateurs qui échangent, discutent entre eux, en interagissant, autour de thématiques qui leur sont chères.

La détection de communautés dépend complètement de la définition qu'on en donne. Elles sont parfois réduites à l'ensemble des gens ayant utilisé un hashtag, ou qui sont amis avec un compte donné. Lorsqu'un graphe dont les nœuds représentent des individus, des algorithmes permettent de détecter des sous-graphes plus denses, mieux connectés, censés représenter des groupes sociaux. Nous suivrons cette piste, en utilisant les messages à la fois comme des arcs, porteur d'interaction entre deux individus, et comme des textes, permettant d'attribuer un label de thématique aux utilisateurs.

Cette démarche permettra d'introduire des scores de cohésion thématique, qui, combinés aux scores topologiques issus de l'analyse des graphes, évaluent la force d'interaction et de partage de centres d'intérêt au sein d'un groupe. En effet, si certains comptes seuls peuvent être considérés comme des influenceurs, il s'avère que la force d'un groupe de petits comptes leur permet d'acquérir parfois une grande visibilité et un impact politique. Nous couvrons ces aspects sous l'intitulé de **détection de communautés thématiques**.

### 1.2.4 Application aux médias sociaux

Nos travaux ne concernent pas toutes les variantes de plate-formes sociales : notamment, nous excluons d'emblée les réseaux sociaux numériques centrés sur les discussions privées, de type « messagerie » (Telegram, Whatsapp), qui sont légalement et techniquement difficiles d'accès. Ainsi, les

14. Plusieurs sources mentionnent cette quantité, dont `http://www.internetlivestats.com/twitter-statistics/`

plate-formes d'échange massif et public (« média social ») de contenus majoritairement textuels sont ciblées : en effet, les traitements nécessaires à l'analyse des contenus multimédia (image, vidéo, son), plus lourds, sont exclus du périmètre de nos travaux.

La plate-forme la plus emblématique correspondant à ces critères est Twitter : de nombreux textes courts y sont largement diffusés, accessibles par tous et notamment par les plus de 300 millions d'utilisateurs actifs, signe d'une bonne vitalité. D'un point de vue purement technique, Twitter permet de collecter automatiquement plus de contenus que Facebook, dont les membres espèrent être lus uniquement par leurs amis.

Malgré ce ciblage sur un type de réseau social, nous souhaitons proposer des contributions valides, qui ne soient pas tributaires d'un unique site Web (aussi gros soit-il). Aussi l'instanciation de notre système n'aura lieu qu'en phase d'évaluation, sur deux jeux de données issus de plate-formes différentes : Twitter et Galaxy2, un petit réseau social basé sur TOR.

## 1.3 Organisation du document

Le triptyque *message, acteur, communauté* est repris dans la suite du document, tout d'abord en partie I, qui propose un état de l'art du domaine. Ainsi le chapitre 2 expose les méthodes connues de calcul du sentiment d'un message, ainsi que les techniques de détermination de l'opinion, de l'émotion et de la posture. Le chapitre 3 détaille les mesures de l'influence des comptes, ainsi que les modèles de caractérisation de leurs comportements. Les algorithmes usuels de détection de communautés, ainsi que les mesures topologiques, calculées à partir de graphes, sont décrits dans le chapitre 4.

La partie II contient nos contributions théoriques, réparties sur les trois volets. Le chapitre 5 se focalise sur la détection de la posture exposée dans des tweets, et introduit une méthode de construction d'une ressource linguistique, les contextosets, afin de désambiguïser les messages et d'améliorer la performance de la détection. Un modèle de caractérisation de l'influence, du comportement et de la position sociale des comptes utilisateurs est proposé en chapitre 6. Enfin, la chaîne de traitement permettant de détecter les communautés d'utilisateurs est décrite en chapitre 7, où sont introduites des mesures de cohésion thématique, qui distinguent les groupes soudés autour d'un thème, ou *communautés thématiques*.

Une évaluation de nos travaux est proposée en partie III. Nos contributions théoriques sont concrétisées et rassemblées en un système, nommé SARTN (Système d'Analyse des Réseaux sociaux sur Trois Niveaux), dont l'implémentation est détaillée en chapitre 8. Ce système est mis en œuvre pour réaliser deux études de cas, en chapitre 9 : la première sur un jeu de données constitué de plus de dix millions de tweets ; la seconde sur l'ensemble des activités effectuées sur un petit réseau social, Galaxy2, qui était disponible sur TOR de 2015 à 2017. Finalement, le chapitre 10 conclut ce document en rappelant nos contributions et leurs limites, et en exposant les perspectives de travaux futurs.


Corto Maltese
*« Je ne suis personne pour juger, je sais seulement que j'éprouve une antipathie innée envers les censeurs. »*

# Première partie

# État de l’art

# CHAPITRE 2

# Analyse du sentiment, détection de l'opinion

L'analyse du sentiment, la détermination de l'émotion, ainsi que la détection de l'opinion ou de la posture d'un texte sont des tâches permettant de calculer l'appréciation émise par les utilisateurs, que ce soit sur des réseaux sociaux, sites de presse, places de marché, envers une marque, entité ou notion. Appliquées à grande échelle, elles permettent de suivre dans le temps des tendances et des évolutions dans le ressenti des clients envers un produit, par exemple. Pour chacune de ces tâches, deux approches sont couramment utilisées pour la mise au point d'un « étiqueteur » : soit l'utilisation de ressources linguistiques et de règles établies *a priori* (« analyseur »), soit l'apprentissage supervisé, à partir d'un corpus annoté (« classifieur »).

La langue elle-même est une difficulté : de nombreuses ressources sont disponibles pour l'anglais, lorsqu'il est correctement rédigé. Les premiers corpus accessibles provenaient d'œuvres littéraires, qui révèlent une langue très différente de la langue parlée ou des usages sur les réseaux sociaux. En conséquence, tant la tâche d'analyse de sentiment que celle de détection d'opinion sont difficiles pour les autres langues, même de grand véhicule (espagnol, français, chinois...), les retranscriptions orales ainsi que les dialectes et sociolectes [Farzindar and Roche, 2013] : ce terme recouvre les spécificités d'usage de la langue sur les réseaux sociaux.

Dans ce chapitre, la section 2.1 définit les notions d'émotion, de sentiment et d'opinion, ainsi que d'une variante à l'opinion : la posture. Les liens entre ces notions y sont discutés, et les particularités des textes issus des réseaux sociaux sont rappelées. Ensuite, nous exposons les chaînes usuelles de l'analyse automatique de texte et les outils nécessaires en section 2.2, puis passons en revue les principales ressources linguistiques permettant le calcul du sentiment en section 2.3, et les techniques d'apprentissage pour le sentiment en section 2.4. Enfin, la section 2.5 se focalise sur les techniques proposées pour la détection d'opinion et de posture. Une synthèse conclut ce chapitre.

## 2.1 Définitions de l'émotion, du sentiment, de l'opinion et de la posture

Dans cette section, nous donnons les définitions de quatre notions, détectées ou extraites du texte, afin d'en clarifier les différences. Les liens entre ces notions sont exposés, et finalement un rappel des difficultés provenant des langages sociaux est proposé.

### 2.1.1 Émotion

L'émotion détectée dans un texte peut représenter soit l'état émotif du locuteur, soit l'impact qu'a le texte sur le lecteur. Le premier aspect permet de suivre tant le ressenti des consommateurs via les critiques produits que la tension interne. L'intérêt et les ouvertures proposées par cet aspect sont liés

au domaine du calcul affectif (*affective computing*) [Picard, 1995]. Le deuxième aspect propose de prévoir les réactions d'une audience, et d'accompagner un message par une charge émotive (il s'agit d'une technique connue de publicité et de propagande).

Deux approches existent pour représenter une émotion informatiquement. Premièrement, par des labels, tels que [Plutchik, 1980] et [Ekman, 1992] les ont proposés, dans deux domaines différents, respectivement la modélisation psychologique théorique et la reconnaissance faciale. Il s'agit, par exemple, de la joie, la peur, l'anxiété, la colère... Les mêmes émotions ont été identifiées par l'analyse d'expressions faciales, montrant une certaine universalité. Il s'agit donc « d'émotions de base » [Ekman, 1992], dont le nombre varie dans la littérature, entre 6 et 8, voire parfois 12, selon le schéma psychologique retenu. Un plus grand nombre d'émotions, incluant les émotions complexes, s'il améliore la qualité du modèle en collant mieux à la réalité, présente toutefois un vrai défi de classification et résulte souvent en des scores (usuellement mesurés par la fiabilité ou le $F_1$) assez médiocres.

Un second mode de représentation se base sur un espace de valeurs possibles : une émotion est un vecteur réel, de deux ou trois dimensions. Plusieurs modèles existent, nous introduisons ici l'un des plus fréquents [Bradley and Lang, 1999]. La première dimension est appelée *valence* : l'émotion est-elle positive comme la joie et la satisfaction, ou négative comme la peur ou la colère ? La seconde dimension, appelée *intensité* [1], repose sur la mesure de l'activité, de l'intensité nerveuse : l'émotion ressentie est-elle intense, comme la colère, ou passive, comme la tristesse ? Une troisième dimension existe parfois : elle représente la *dominance*, la résistance de la personne à ses émotions, allant de l'émotion contenue jusqu'à la perte de contrôle émotionnelle.

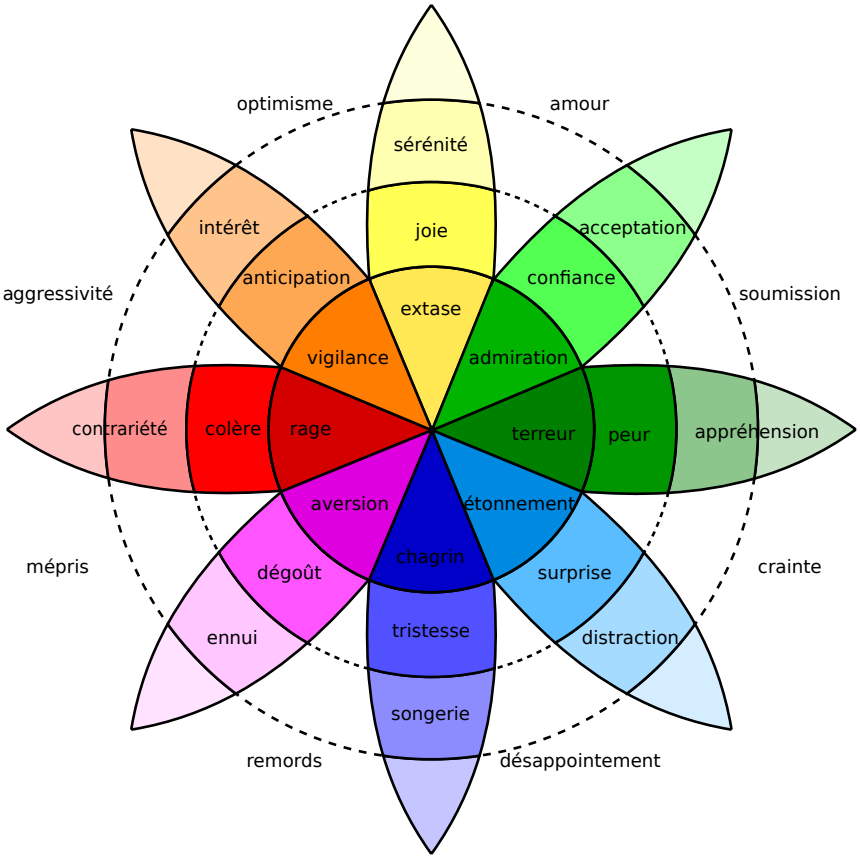



FIGURE 2.1 – La roue des émotions de Plutchik, selon l'encyclopédie Wikipédia

1. ou *arousal* en anglais

Reprenant les labels, le schéma [2] représenté en figure 2.1 propose différents niveaux d'intensité (sur l'axe radial) des émotions, et les place selon un certain ordre qui suggère une « proximité » entre émotions. Il illustre ainsi une correspondance entre la première approche, par étiquettes figées, et la seconde par l'espace de valeurs.

Dans ce domaine très actif de la détection d'émotions, [Yassine and Hajj, 2010] a proposé un modèle pour détecter et exploiter les émotions ; ce modèle a été appliqué à un corpus issu de Facebook. Ces travaux fonctionnent à partir de postulats parfois forts : par exemple, il semblerait que les textes ne contenant pas d'émotion seraient objectifs, et donc sans opinion. Ce module permet d'ouvrir ensuite de nouvelles perspectives à l'analyse des réseaux sociaux, avec par exemple la détection d'amitié entre utilisateurs selon l'intensité émotive de leurs échanges [Yassine and Hajj, 2010].

Considérant qu'il existe une valeur émotionnelle intrinsèque aux mots, [Mohammad and Turney, 2010] a créé un dictionnaire multi-lingue, appelé « emotion-lexicon » (EmoLex), grâce à de l'annotation participative. Un modèle à 8 labels a été utilisé pour modéliser les émotions : à chaque couple terme-émotion correspond un booléen, indiquant si le terme est porteur de l'émotion en question.

[Munezero et al., 2015] a utilisé un tel dictionnaire pour attribuer des labels émotifs à des tweets. Pour consolider les prédictions liées à chaque message, il a procédé à l'agrégation de plusieurs tweets d'un même utilisateur, selon l'intuition que « l'émotion est un sentiment de long terme » : un auteur conserverait son émotion, qui évoluerait sans rupture au fil du temps. Cependant, les émotions sont souvent associées aux sautes d'humeur, ce qui fragilise cette hypothèse.

### 2.1.2 Sentiment

Le sentiment d'un texte, ou polarité, c'est à quel point un texte est mélioratif ou péjoratif. Dans cette situation, le sujet importe peu ; est-ce que la tonalité générale est positive ou négative ? Cette définition souffre de quelques *a priori* mal définis : il faut déjà que le texte soit porteur d'une subjectivité. S'il est purement informatif, il n'y a guère de place pour le sentiment. D'autre part, il ne s'agit pas d'aller jusqu'à détecter l'état émotif de l'auteur lors de la rédaction, mais de s'arrêter au texte. Une autre piste d'interprétation du sentiment s'intéresse au domaine de l'attitude et de la posture. Est-ce que le sentiment d'un texte correspond à la posture de son auteur envers le sujet du texte ?

Le sentiment se focalise sur le ton du texte, sans se préoccuper de la thématique. Il est souvent représenté par un nombre réel, par exemple dans $[-1, 1]$, où $-1$ représente un texte très négatif, et 1 très positif. Une excellente revue de ce domaine a été produite par [Pang et al., 2002], qui se penchait sur les prédictions du score attribué par les auteurs de critiques (score mesuré en nombre d'étoiles, souvent de 1 à 5) à partir des textes des commentaires sur des sites de revues de films.

### 2.1.3 Opinion

L'opinion est habituellement représentée par un tuple qui correspond au mieux à un texte. Ce tuple contient 5 éléments : l'auteur, la date d'expression, la cible du texte, l'aspect de la cible, et la polarité du texte (ou sentiment) [Pang and Lee, 2008]. La notion d'opinion avait été introduite plus tôt, par l'exploitation des textes de revues de produits et le calcul du nombre d'étoiles associé [Pang et al., 2002], de 1 (client mécontent) à 5 (client satisfait).

Autrement formulée, la détection d'opinion [3] consiste à déterminer chacune des informations suivantes, constituantes de l'opinion [Liu, 2010] :

— **Auteur :** qui parle ? Qui est l'émetteur de l'opinion ?

— **Date :** quand ? À quel moment est exprimée l'opinion ?

— **Cible :** de quoi parle le texte ? Sur quoi porte l'opinion exprimée ?

— **Aspect de la cible :** de quelle caractéristique ou attribut de la cible parle le texte ?

2. Issu de : `https://fr.wikipedia.org/wiki/Émotion`
3. parfois nommée *opinion mining*

— **Polarité :** comment (bien, ou mal ?) parle-t-on de la cible ? L'auteur est-il favorable ou non à la cible du message ?

Le plus souvent, l'auteur et la date peuvent être obtenus via les méta-données, bien qu'il y ait du discours rapporté sur les médias sociaux. La cible et la polarité sont plus compliquées à calculer ou extraire à partir du texte.

### 2.1.4 Posture, attitude

Bien souvent, la détection de toutes les opinions n'est pas nécessaire ; elle peut consister à seulement surveiller les opinions vis-à-vis d'un nombre restreint de cibles. Ainsi, il devient intéressant de développer un module spécifique à la thématique d'intérêt.

Lorsqu'on se focalise sur une cible, on parle de posture vis-à-vis de la cible, ou d'attitude, ou de position. En anglais, cela s'appelle *stance detection* : il s'agit de déterminer la polarité du texte (par un label : en faveur, neutre ou opposé ; ou par une valeur, entre -1 et 1).

Une application notable de la détection de posture est basée sur l'analyse des publications sur un site de débats, *ConvinceMe.Net* [Anand et al., 2011]. L'une des difficultés majeures signalées provient des réparties : des réponses, dépendant complètement du message précédent pour être interprétées et comprises. Ces réparties, ne portant parfois aucune mention d'entité, peuvent être attribuées à n'importe quel attitude ou camp si elles ne sont pas mises en relation avec l'ensemble du dialogue.

Une tâche d'évaluation internationale nommée SemEval [4] a proposé un corpus pour formaliser ce problème et améliorer l'état de l'art mondial. C'est une tâche difficile, car ce corpus, portant sur six thématiques, contient des tweets mentionnant diverses cibles. Par exemple, sur la thématique « Donald Trump » se trouvent des tweets mentionnant des événements, soutiens et opposants à sa candidature, de polarités variables. Ainsi, une attitude de soutien envers un événement d'opposition au candidat risque de résulter en une mauvaise prédiction. Poser le problème en détection de l'attitude ne dispense pas de déterminer ou au minimum de prendre en compte la cible du texte.

### 2.1.5 Liens entre émotion, sentiment, opinion et posture

Intuitivement, les liens entre sentiment et opinion sont forts : un sentiment positif envers une entité implique une opinion positive envers cette même entité. De même, la posture peut être vue de manière simplifiée comme l'opinion envers une unique entité cible. Cependant, les cas d'application varient et débouchent, dans la pratique, sur des résultats à la fiabilité variable. Le sentiment n'est pas aisé à exploiter selon les applications, et ni l'opinion ni la posture ne peuvent être extraites systématiquement de n'importe quel texte à l'heure actuelle, au vu des performances atteintes sur des tâches comme SemEval [Mohammad et al., 2016].

Les quelques éléments de comparaison proposés en table 2.1 alimentent la discussion suivante. Le sentiment et l'émotion, génériques et mesurables sur tout type de texte car intrinsèques aux textes, peuvent être calculés seulement sur la polarité des mots utilisés grâce à un dictionnaire de sentiment ou d'émotion, ou par l'apprentissage sur un corpus. Il en résulte parfois une information contre-intuitive. Ainsi, une revue négative d'un produit peut être exprimée avec des mots favorables, positifs : *« un film gentil, déclaratif et convenu »*, et l'opinion (négative) différera du sentiment (positif, du fait de « gentil »). Cependant, la présence du sentiment dans tous les textes permet de générer des statistiques pour suivre, par exemple, le sentiment global sur un flot de nouveaux contenus ; de son côté, l'opinion, plus fine et précise, nécessite une connaissance implicite pour faire le lien entre les entités mentionnées, et les entités analysées.

L'opinion ajoute notamment le sujet du texte. À première vue, le sujet grammatical pourrait suffire pour savoir sur quoi porte l'opinion, par exemple dans *« ce film est excellent »*. Cette intuition ne permet malheureusement pas de bien reconnaître les opinions, lorsque le sujet de la phrase est l'auteur du texte par exemple : *« j'ai beaucoup apprécié ce film »*. Une phase de reconnaissance des

4. http ://alt.qcri.org/semeval2016/task6/

TABLEAU 2.1 – Élements de comparaison entre sentiment, opinion et posture

| | Sentiment et Émotion | Opinion | Posture |
|---|---|---|---|
| Cible | non exprimée | thème du texte | cible de l'analyse |
| Pré-requis du système | dictionnaire, ou corpus | connaissance implicite, grammaire, accord annotateurs | système *ad hoc* |
| Domaine d'application | tout texte | revue de produits, de films | textes politiques |

entités nommées présentes améliore la gestion de cette détection de la cible des opinions. Parfois, une connaissance préalable de la cible permet de mieux la reconnaître lorsque qu'elle n'est pas explicitement mentionnée, mais seulement suggérée.

La détection de posture fixe une fois pour toutes la cible des polarités à calculer : la lecture des textes ne se fait plus qu'au prisme de la cible choisie. Il en résulte la nécessité d'avoir un système, ou au moins un modèle, figé pour le besoin courant. Tous les textes sont soit pour, contre ou sans opinion tranchée envers cette cible... Une catégorie « hors sujet » peut permettre d'écarter les textes sur lesquels le modèle ne peut pas s'appliquer. De plus, la limite est floue entre les textes interprétables, pour lesquels il est certain qu'une posture est présente. Dans le cadre d'un besoin politique, s'il faut analyser la posture de tweets envers un candidat, comment faut-il interpréter les messages parlant d'un meeting ou d'un député, sans savoir qui le meeting ou le député soutiennent ?

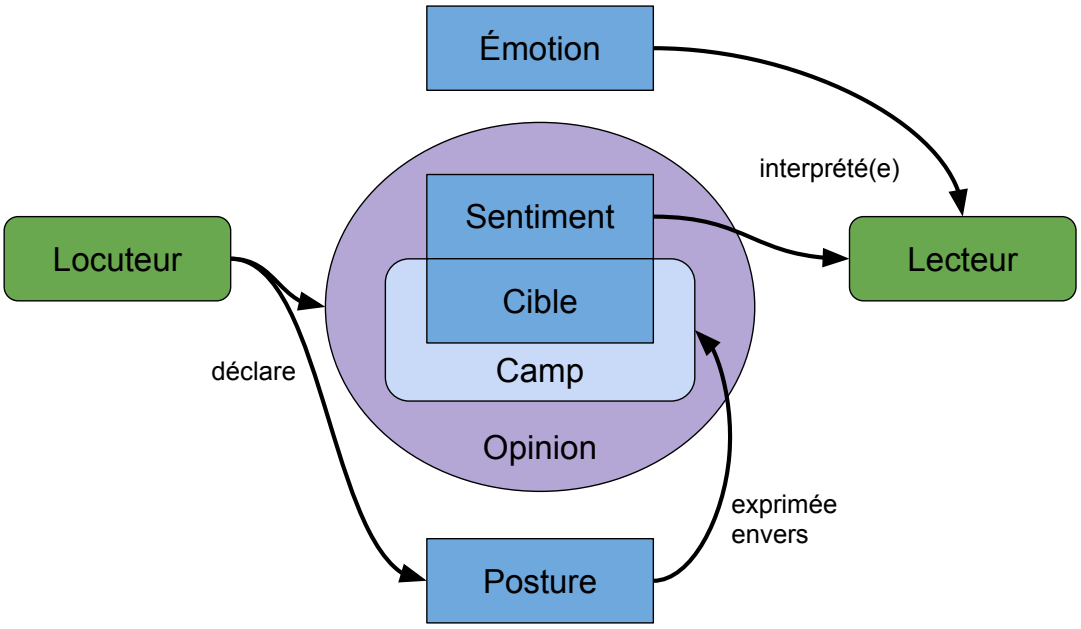



FIGURE 2.2 – Notions, élocution et lecture

Pour clarifier les relations entre ces notions, le locuteur et le lecteur du message, la figure 2.2 récapitule les grands axes des définitions introduites. Le locuteur, auteur du message, déclare son opinion (et sa posture). Cette opinion est composée d'un sentiment (ou polarité) et d'une cible. La cible fait référence à un ensemble d'entités, le *camp*, au regard duquel la posture est exprimée. Les valeurs de sentiment et d'émotion correspondent à ce que le lecteur interprète, écartant ainsi la personnalité de l'auteur du texte. Notons que le locuteur d'un message se transforme en lecteur face aux publications des autres : il s'agit toujours de « comptes utilisateurs ».

La détermination de ces notions présente des avantages et inconvénients spécifiques. Alors que la détection de sentiment ne permet pas facilement de répondre à un besoin spécifique, l'approche de détection de posture requiert beaucoup plus d'informations. En outre, il faut maintenir les ressources à jour : le vocabulaire politique évolue rapidement, chaque nouveau thème de débat mettant à jour les discours et phrases-types de chacun ; de même en marketing, les nouveaux produits, tendances et marchés doivent être suivis, ce qui nécessite de créer constamment de nouvelles vérités-terrain, là où la détection de sentiment est finalement assez stable dans le temps.

Malgré tout, sur ces tâches de détection d'opinion, de sentiment ou de posture, la validation repose sur une annotation des textes par des êtres humains. Il faut alors vérifier en permanence la qualité de l'annotation et sa validité, par exemple grâce au calcul d'un score inter-annotateurs. Il est fréquent d'avoir des scores bas lorsque la tâche est difficile comme par exemple la détection de l'ironie, ou que les textes sont intrinsèquement difficiles.

### 2.1.6 Les caractéristiques particulières des *sociolectes*

Ce qui était autrefois nommé « fautes de frappe », liées à la difficile manipulation de petits claviers (de téléphone, ou tactiles) est devenu un langage, qui évolue rapidement dans la population [Farzindar and Roche, 2013]. Rallongement des voyelles dans les mots, usage intensif de la ponctuation, émoticônes et mots-dièses sont les symboles de cette évolution, qui apparaît dans de nombreuses langues. [Maynard et al., 2012] invoque d'autres difficultés dans l'annotation des tweets, et suggère que les systèmes devraient d'abord évaluer la pertinence, l'opportunité d'analyser un tweet avant d'en identifier la cible. Ce n'est que dans un second temps que des points tels que le sarcasme ou la négation peuvent être considérés pour établir la polarité d'un message. Le sarcasme n'est pas spécifique aux sociolectes, mais est tout de même notoirement fréquent, conjointement avec d'autres formes d'humour, et demeure difficile à détecter.

La figure 2.3 présente un exemple léger des styles d'écriture sur les réseaux sociaux. Ici, un tweet institutionnel, émis par le CEA, utilise des mots-dièses, émojis, un bon nombre de mentions, et surtout un GIF. Chacun de ces éléments a une fonction dans le texte : la première émoji sert à rendre positif un texte initialement peu avenant, la seconde à introduire le lien web. Le GIF attire l'attention plus efficacement que le message lui-même, d'autant plus qu'il contient un chat. Enfin, les sept mentions d'utilisateur sont destinées à faciliter la propagation du tweet, qui sera vraisemblablement mis en avant par les partenaires mentionnés.

Dans cette section, nous avons présenté les quatre notions d'émotion, de sentiment, d'opinion et de posture, leurs liens et différences, et nous avons rappelé les difficultés provenant des textes issus des réseaux sociaux. La section suivante présente la chaîne de traitement usuelle pour détecter les notions recherchées, et éventuellement résoudre le problème posé.


FIGURE 2.3 – Exemple de tweet

## 2.2 Chaîne de traitement usuelle

L'immense majorité des traitements sur une chaîne de caractères nécessite une étape préliminaire de nettoyage, de préparation. Le texte en entrée, écrit par un utilisateur, sans contraintes, ne correspond pas toujours à un format spécifique. On parle de « contenu généré par l'utilisateur » [5].

La figure 2.4 propose une vue générale d'une chaîne de traitement usuelle, préparant le texte afin d'obtenir une représentation vectorielle (souvent un vecteur $tf.idf$), qui est l'un des formats favoris pour alimenter un étiqueteur. L'étiqueteur est le module générique qui effectue une prédiction, que ce soit la détermination de la polarité de l'opinion, de l'émotion, du sentiment ou de la posture : il peut s'agir d'un classifieur statistique, ou d'autre chose. De même, d'autres représentations que le $tf.idf$ existent ; nous y revenons plus tard dans cette section.

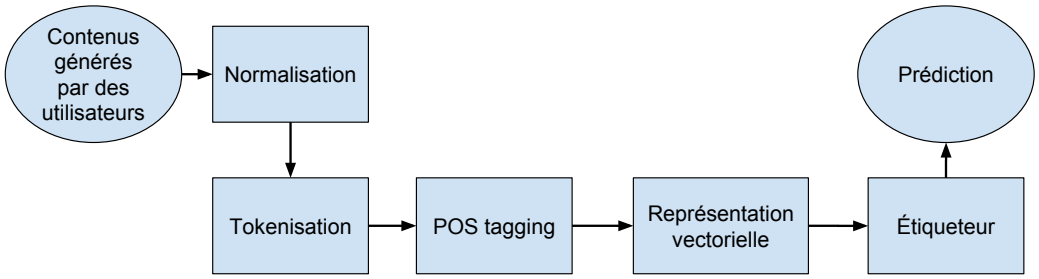



FIGURE 2.4 – Vue globale d'une chaîne d'analyse de texte

Dans cette figure 2.4, le texte est d'abord ingéré par le module de Normalisation, qui permet de nettoyer la donnée d'entrée. La chaîne de caractères est ensuite transmise au module de Tokenisation, qui découpe la chaîne en termes. L'information est transmise sous forme de liste de termes au module suivant, le POS tagger (*Part of speech*) : le plus souvent à base de règles, celui-ci détermine le rôle grammatical de chaque terme (nom, adjectif, interjection...) dans la phrase. Ainsi, un terme-adjectif sera considéré différemment d'un terme-nom [6]. Ensuite, on procède à la transformation en vecteur, ici en $tf.idf$. Finalement, cette représentation du texte est ingérée par un étiqueteur, qui détermine, souvent à l'aide de ressources linguistiques, le label recherché. Tous ces modules ne sont pas requis ; un choix peut être fait, par exemple de limiter les corrections apportées par la Normalisation pour conserver la forme originale du texte.

### 2.2.1 Normalisation

L'objectif de ce module est de modifier légèrement le texte de façon à faciliter son traitement statistique, de corriger rapidement les orthographes inventives (doublement ou triplement de quelques voyelles, correction orthographique, ...). Par exemple, il est courant de passer tous les caractères en minuscules, et de corriger les erreurs de frappe les plus fréquentes. En anglais, « *can't* », « *cannot* » et « *cant* » sont uniformisés en un même mot. Les caractères spéciaux et la ponctuation non-structurante sont retirés, excluant les émoticônes, les flèches, les lignes, et remplaçant d'éventuels points de suspension par des points.

La normalisation a un impact sur la densité des occurrences des termes : trop de corrections résultent dans l'uniformisation entre des termes différents. Pas assez de corrections, et la taille du vocabulaire explose (en français, *non*, *noon*, *nooon* et *nooooooooon* seraient plusieurs mots différents).

5. ou *user-generated content*.
6. par exemple, le terme « objectif »

### 2.2.2 Tokenisation

La tokenisation assure la transformation d'une chaîne de caractères en phrases composées de mots ou termes, le plus souvent en coupant sur les espaces et la ponctuation. Intuitivement, une phrase finit par un point, et deux mots sont séparés par un espace.

Cette phase peut être plus ou moins sophistiquée et considérer l'ensemble de la ponctuation ainsi que les propositions subordonnées : une proposition est transformée en phrase, où le pronom relatif est remplacé par ce qu'il réfère. La tokenisation est très dépendante de la langue et du type de documents : un tweet ne se découpe souvent qu'en une seule phrase ; en chinois, les mots ne sont pas séparés par des espaces.

### 2.2.3 Analyse morpho-syntaxique et lemmatisation

Aussi nommée Étiquetage des parties de discours, ou *Part-of-Speech (POS) tagger* en anglais, cette phase réalise l'annotation des rôles grammaticaux. La forme initiale du mot est souvent requise, car les majuscules permettent de suggérer des noms propres, et les formes fléchies (accords des adjectifs, conjugaisons, déclinaisons...) aident à déterminer le rôle de chaque terme.

Cette phase est souvent réalisée conjointement à la lemmatisation, qui consiste à regrouper sous le même mot ses formes fléchies : par exemple, *« better »* est le superlatif du lemme *« good »*. À la place de la lemmatisation, on utilise parfois la racinisation, ou *stemming*, qui retire les suffixes identifiés comme tels.

L'un des outils les plus répandus, qui fonctionne bien sur les textes corrects, s'appelle TreeTagger et fonctionne à base de règles et de ressources linguistiques (dictionnaires, tables de verbes...) [Schmid, 1994].

Un outil adapté aux tweets en anglais, basé sur l'enrichissement des ressources classiques via des clusters de Brown (fusion de mots proches), a été réalisé par [Owoputi et al., 2013]. Il propose une annotation (similaire à celle obtenue par TreeTagger, qui recourt à la base PennTreeBank) des tweets. Cependant ce modèle est confronté à des abbréviations / sociolectes / smileys n'apparaissant qu'une seule fois dans les corpus, et ne réagit pas toujours convenablement (par exemple, lemmatisation de noms de famille et de sigles).

### 2.2.4 Projection du texte vers un autre espace

La représentation informatique d'un texte nécessite une étape de projection de la chaîne de caractères, reçue en entrée, vers une liste de caractéristiques, permettant la construction d'un vecteur observation. L'impact des caractéristiques choisies est très important, comme montré dans diverses études, telle [Tan et al., 2002] qui compare l'usage d'unigrammes et bigrammes de mots. Nous présentons ici les caractéristiques courantes dans le domaine :

- **représentation des termes.** Les caractéristiques sont ici construites à partir des mots :
  - unigrammes. Il s'agit du terme lui-même.
  - bigrammes. Chaque association d'un mot et du mot suivant est comptée. Le mot seul n'est pas pris en compte, seulement comme partie d'un couple. *« Hello, my dear »* contient les bigrammes (*hello*, *my*) et (*my*, *dear*).
  - n-grammes. Le même principe est appliqué aux chaînes de $n$ mots. Le choix de $n$ varie selon la langue, et reste souvent faible (inférieur à 5), afin de limiter la taille du vocabulaire, et d'augmenter la fréquence des motifs représentés.
- **n-grammes de caractères.** Chaque suite de $n$ lettres (ou symboles) est une caractéristique. Par exemple, avec $n = 3$, *« Hello »* contient (H,e,l), (e,l,l), (l, l, o). Ces n-grammes de caractères permettent de considérer les orthographes inventives et sociolectes. En effet, les signes tels que « :) » ou « ! ? » sont très présents parmi les tweets. Proposés pour introduire une représentation du texte indépendante de la langue, les n-grammes de caractères sont souvent utilisés pour classifier un texte selon sa langue [Damashek, 1995].

— **word embeddings**
  — **des mots.** Une méthode récente a introduit le concept des « word embeddings », via la méthode **word2vec** [Mikolov et al., 2013]. Par apprentissage, chaque mot est placé dans un espace vectoriel de grande dimension (souvent 200 ou 500 dimensions) à partir de leurs co-occurrences. Il en résulte des clusters parfois sémantiques, parfois syntaxiques aux propriétés intéressantes : par exemple, il existe une direction des superlatifs (« *biggest* », « *shortest* », ... sont placés dans une même direction). Une autre méthode basée sur des réseaux neuronaux convolutifs répond au même défi [Kalchbrenner et al., 2014].
  — **des phrases.** Une amélioration du modèle *word2vec* a permis de représenter des phrases dans cet espace, plutôt que des mots isolés [Le and Mikolov, 2014]. Une première version, *phrase2vec*, permet de représenter des couples de mots (tels *Los Angeles*, ou *New York*), et une seconde version, *doc2vec*, projette des documents complets, ce qui nécessite un corpus d'apprentissage de très grande taille (plusieurs dizaines de millions de documents, voire milliards).

### 2.2.5 Bag of Words, fréquences et tf.idf

Pour utiliser des algorithmes d'apprentissage, il est nécessaire de construire des vecteurs à partir du texte. L'une des façons, c'est d'y représenter des n-grammes (unigrammes, bigrammes, ...) de mots ou de caractères. Ces vecteurs prennent plusieurs formes :

— **vecteurs binaires de présence.** Pour chacune des caractéristiques retenues, un booléen marque sa présence ;
— **vecteurs de comptage.** Un entier compte le nombre d'occurrences de la caractéristique ;
— **vecteurs de fréquence.** Un réel représente la fréquence d'occurrence de la caractéristique ;
— **vecteurs tf.idf.** Il s'agit du produit de deux fréquences. $tf$ représente la fréquence d'occurrence de la caractéristique dans le document, et $idf$ le nombre de documents la contenant dans le corpus.

Comme les vecteurs peuvent être de très grande dimension si toutes les caractéristiques étaient retenues, il est courant de ne garder que les plus fréquentes ou les plus informatives. Des dimensions typiques sont 2 000, 5 000 ou 10 000 caractéristiques. Le stockage est réalisé dans des matrices creuses. Ces approches sont appelées en « sac de mots » (Bag of Words), car elles ne préservent pas l'ordre des mots ou caractéristiques dans la phrase.

Dans cette section, la chaîne de traitement usuelle a été présentée, ainsi que les principaux composants du pré-traitement sur les textes. La tâche accomplie par le module d'*Étiqueteur*, maillon présenté en figure 2.4, prend plusieurs formes, dont la première à être présentée est l'analyse du sentiment.

## 2.3 Analyse du sentiment basée sur des ressources

Souvent focalisée sur les revues de produits, l'analyse de sentiment a connu, dans un premier temps, une période de construction de ressources linguistiques : listes d'émoticônes positives et négatives, d'adjectifs et autres termes. Une liste des principales ressources pour la langue anglaise est proposée dans cette section. Dans un second temps, ces ressources sont exploitées, parfois par des règles simples, parfois par des systèmes plus élaborés. Enfin, des approches basées sur un corpus d'apprentissage annoté pour déterminer le sentiment sont présentées.

### 2.3.1 Principe

Étant donnée une liste de termes provenant d'un texte, cette approche utilise des ressources qui attribuent des polarités à chacun des mots ; une somme pondérée permet ensuite de calculer le sentiment global du texte analysé. La figure 2.5 propose une telle chaîne « naïve » : à partir d'une phrase

ou document dont il faut déterminer la valence, une liste de mots valués est extraite. En effet, de nombreux mots ne portent pas d'information affective. Parallèlement, des informations syntactiques sont extraites (typiquement, la détection de négation ou de propositions subordonnées conditionnelles). Ensuite, une étape de combinaison des scores des mots valués permet de produire un score de sentiment. Cette phase de combinaison est parfois assez simple, telle la moyenne des valences de mots ; elle peut aussi être plus complexe, considérant que les adverbes multiplient la valence des adjectifs, ou que la négation transforme une valence positive en négative.

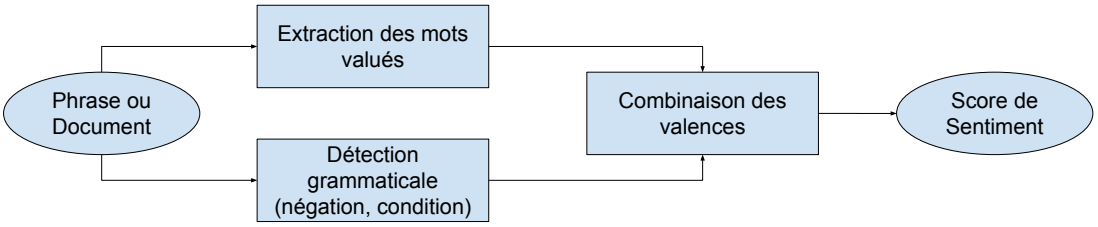



FIGURE 2.5 – Calcul classique du sentiment d'un document

### 2.3.2 Les principales ressources linguistiques en anglais

La table 2.2 liste les ressources linguistiques majeures disponibles pour l'anglais, langue sur laquelle la majorité des travaux se focalise. Il est courant de trouver des traductions de ces ressources pour d'autres langues. Trois de ces ressources contiennent certes des informations sur l'émotion (LIWC, ANEW et EmoLex), mais aussi sur le sentiment. Nous revenons plus en détail sur chacune de ces ressources, une par une.

TABLEAU 2.2 – Liste des principales ressources linguistiques pour l'analyse de sentiment

| Ressource | Nom Complet | Référence | Commentaires |
|---|---|---|---|
| WNA | WordNetAffect | [Strapparava et al., 2004] | petit mais expert |
| SWN | SentiWordNet | [Baccianella et al., 2010] | large mais bruitée |
| LIWC | Linguistic Inquiry and Word Count | [Pennebaker et al., 2001] | style écriture |
| ANEW | Affective Norms for English Words | [Bradley and Lang, 1999] | petit mais expert |
| EmoLex | Emotion Lexicon | [Mohammad and Turney, 2010] | émotions |
| MPQA | Question-Answering | [Wilson et al., 2005] | attitude, par mot |

**WordNetAffect :** WordNet [Miller, 1995] est un dictionnaire de synonymes en anglais, compilé par des linguistes. Il a bénéficié de plusieurs extensions dont WordNet-Affect [Strapparava et al., 2004], qui attribue des valences de sentiment à un faible nombre d'ensembles de synonymes (environ 1 900 termes). Cette attribution a été effectuée par une équipe de linguistes, et est considérée comme juste.

**SentiWordNet :** Pour élaborer SentiWordNet [Baccianella et al., 2010], ses auteurs sont repartis du dictionnaire WordNet, en utilisant un mécanisme de propagation des positivité, neutralité et négativité des mots le long des relations entre termes, contenues dans WordNet. Cette propagation a pour source une poignée de termes « intrinsèquement positifs » (resp, négatifs), et s'accompagne de quelques mesures de fiabilités pour s'assurer de la validité des scores de valence des mots. Sa précision est globalement moins bonne que WordNetAffect, mais est compensée par la taille de son vocabulaire [Şerban et al., 2012], qui améliore la couverture, donc le rappel.

Avec la même philosophie, [Asghar, 2014] a créé un dictionnaire affectif de termes de *slang* (argot anglais) : en utilisant un corpus de messages issus de réseaux sociaux, il attribue aux termes d'argot une polarité calculée à partir des polarités des messages qui les contiennent.

**LIWC :** Linguistic Inquiry and Word Count, abrégé LIWC (se prononce comme « *Luke* ») est à la fois la ressource linguistique, et le programme qui l'exploite. Un premier niveau regroupe les termes en catégories : usage du conditionnel, de la première personne du singulier ou du pluriel, charge émotive, ... Les aspects pris en compte sont nombreux. Ces catégories peuvent ensuite servir à calculer des indicateurs de plus haut niveau comme la « confiance en soi » ou « l'authenticité » montrée par l'auteur dans son texte [Pennebaker et al., 2001]. Une première version, datant de 2001, a été traduite en allemand et en espagnol. La version anglaise a été mise à jour en 2007.

**ANEW :** ANEW regroupe 600 mots anglais, dont l'annotation a été réalisée en demandant à des individus de donner des scores pour chacun des mots de manière isolée. Le score, représentant les émotions, est tri-dimensionnel : *valence* (de négatif à positif), *arousal* ou intensité (de faible à forte), et *dominance* (émotion contenue, ou hors de contrôle) [Bradley and Lang, 1999]. L'utilisation de la première dimension, la *valence*, permet déjà de calculer une polarité de sentiment.

**Emolex :** EmoLex est un dictionnaire associant des ensembles de synonymes (ceux de WordNet) à des labels d'émotions : un booléen indique si le terme est associé à l'une des 8 émotions du modèle de [Plutchik, 1980]. De même, ce dictionnaire indique si le mot est positif, ou négatif. Cette annotation a été réalisée par crowdsourcing sur la plateforme Amazon Mechanical Turk [Mohammad and Turney, 2010].

**MPQA :** La ressource MPQA originale [Wiebe et al., 2005] est composée de documents annotés : pour chaque énoncé, qui est l'agent émetteur de l'expression, quelle attitude est exprimée, et envers quelle entité cette attitude est dirigée. [Wilson et al., 2005] a ajouté une annotation « d'expression subjective » sur le corpus MPQA. Il distingue les polarités « a priori », telles celles proposées par SWN, des polarités contextuelles : la valence d'un mot est modifiée par le texte dans lequel il apparaît. Une troisième version introduit une annotation spécifique vers des entités « événements » [Deng and Wiebe, 2015], sur un ensemble de 70 documents exprimant 1 287 attitudes envers 1 213 entités.

### 2.3.3 Exploitation de ressources par des règles

Pour profiter des forces de plusieurs ressources, [Khan et al., 2014] a combiné un classifieur qui prend en compte les émoticônes, le sentiment calculé à l'aide d'un dictionnaire *ad hoc*, et le sentiment basé sur SentiWordNet. Le chaînage de ces briques est présenté en figure 2.6 : le choix porte d'abord sur les émoticônes, qui peuvent suffire à attribuer un label (positif, ou négatif), puis sur les deux autres modules. Seul le dernier module peut attribuer le label neutre. Selon [Khan et al., 2014], l'étiqueteur par émoticônes (enhanced emoticon classifier) résulte en une excellente précision, mais un faible rappel (environ 17%) : tous les tweets ne contiennent pas de smiley explicite. Les résultats des trois modules combinés sont meilleurs qu'isolés : grâce à l'ordre dans lequel les modules s'expriment, une fiabilité de 86% est atteinte sur un ensemble de 6 corpus de tweets, regroupant un total de 2 116 messages, annotés via crowdsourcing.

Le système Vader, pour *Valence Aware Dictionary for sEntiment Reasoning* [Hutto and Gilbert, 2014], repose sur un analyseur basé sur des règles, compilant LIWC, ANEW et SWN : une liste de 7 500 couples mot-valence en sont extraits. Quelques règles grammaticales permettent une annotation précise, validée empiriquement sur quatre corpus : tweets, revues de produits, de films, et articles de presse. Ce module propose 4 scores réels, de positivité, de négativité, d'objectivité, et enfin un score composé correspondant à la polarité de sentiment, exprimé entre $-1$ et 1. Les auteurs concèdent que l'apprentissage automatique rend de bien meilleurs résultats sur certains types de texte, ici sur les revues de film. Ce module fait désormais partie du package Python NLTK.

Enfin, un autre système à base de règles grammaticales, *Opinion Observer* [Ding et al., 2008], se démarque par la finesse de son algorithme : d'une part, les ressources linguistiques sont enrichies, notamment avec des expressions idiomatiques ; d'autre part, un traitement spécifique est appliqué aux

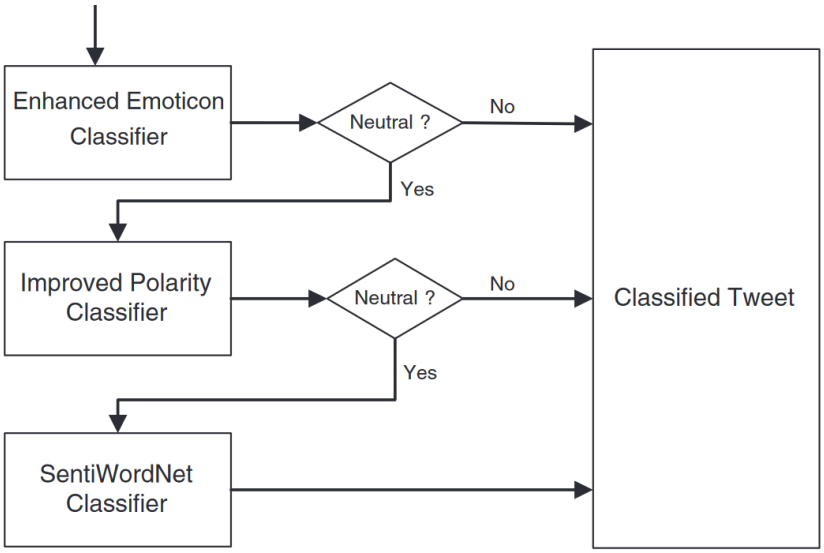



FIGURE 2.6 – Étiqueteur de sentiment proposé par [Khan et al., 2014] (extrait de la Fig. 3, p. 248)

phrases contenant « mais », en parallèle du traitement réservé à la négation. Sur un ensemble de 445 revues de 8 produits (appareils photo, lecteurs MP3 et DVD) différents, ce système atteint des $F_1$ (moyenne harmonique de la précision et du rappel) allant de 0,83 à 0,96.

Ces approches requièrent à la fois des exemples d'expression d'opinions, souvent restreints à un domaine assez limité (soit la critique de film, soit la critique de produits). De plus, un linguiste est nécessaire pour mettre au point des règles pertinentes, et souvent nombreuses. L'avantage de cette approche repose dans la qualité, souvent bonne, de l'analyse lorsqu'elle est réalisée dans le bon domaine. Dans la section suivante, nous présentons des techniques plus statistiques, qui évitent la phase d'analyse linguistique.

## 2.4 Analyse du sentiment basée sur les approches statistiques

Dans cette section, nous présentons des techniques recourant à l'apprentissage automatique : d'abord les techniques se basant uniquement sur un corpus d'apprentissage, puis dans un second temps, des techniques hybrides qui exploitent aussi des dictionnaires de sentiment ou d'autres ressources créées par des experts.

### 2.4.1 Catégorisation par apprentissage sur un corpus

La classification du sentiment par apprentissage s'alimente des techniques provenant de la catégorisation automatique de texte : à l'origine, il s'agissait par exemple de déterminer les thématiques principales d'un corpus [Pang and Lee, 2008]. Chaque document reçoit un label parmi quelques catégories prédéfinies (sport, culture, économie... pour des articles de presse, par exemple). L'utilisation de classifieurs tels que Bayes naïf ou SVM [7] sur des unigrammes de mots est présentée comme facile à adapter à un nouveau corpus [Pang and Lee, 2008, Liu, 2010].

Les contenus dits sociaux, générés par les utilisateurs, sont très spécifiques à leur plateforme d'émission. L'émergence des orthographes inventives a mené [Pak and Paroubek, 2010] à rassembler un corpus pour l'analyse de sentiment, spécialement constitué de tweets. La taxonomie est la suivante : les tweets sont d'abord soit objectifs (donc neutres), soit subjectifs, auquel cas une règle de présence d'émoticônes positives ou négatives a permis d'attribuer une étiquette de sentiment. Un

7. Support Vector Machine

corpus équilibré de 300 000 tweets est ainsi constitué. Dans un second temps, une chaîne usuelle de classification est appliquée (nettoyage, projection en n-grammes, utilisation d'un Bayes naïf), et validée sur un ensemble de 216 tweets annotés à la main.

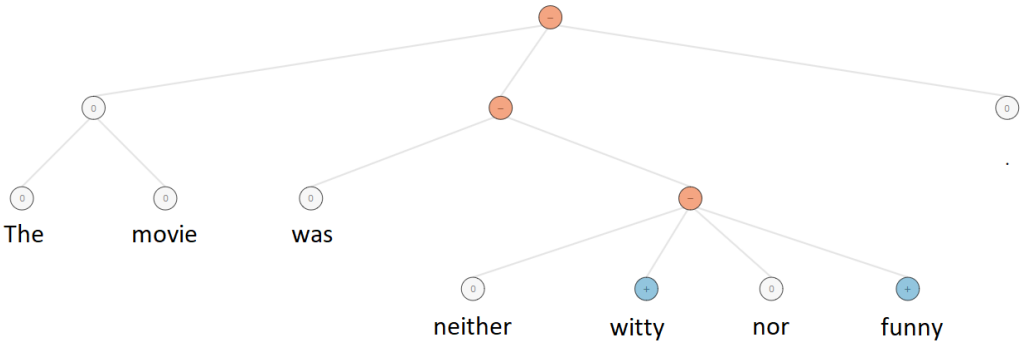



FIGURE 2.7 – Pondération des mots dans une phrase ; issu de Stanford NLP [Socher et al., 2013]

[Socher et al., 2013] propose un modèle qui s'alimente d'arbres représentant la phrase (*Sentiment TreeBank*), ce qui permet de pondérer le poids des mots et de considérer automatiquement la négation. La figure 2.7 prend en exemple la phrase « *This movie was neither funny nor witty* », et montre comment la méthode, constituée d'un RNN [8], parvient à reconstruire la polarité de la phrase. Une démonstration en ligne est disponible [9], mais rencontre des difficultés sur les textes incorrects.

En 2013, sur une tâche de calcul de polarité du sentiment au niveau du message lors de SemEval2013, les meilleures méthodes recouraient à SentiWordNet ou à MPQA ; 8 utilisaient un SVM, et 7 un Bayes naïf. Parmi les 28 systèmes proposés, se trouvaient encore des classifieurs linéaires, des classifieurs à base de règles, mais aussi des mélanges de plusieurs modèles. Les meilleurs systèmes ont atteint une $F_1 = 0.69$ (subtask-B : sentiment du message global) [Nakov et al., 2013] : la tâche est difficile.

### 2.4.2 Méthodes hybrides

[Bahrainian and Dengel, 2013] propose de calculer une polarité de sentiment grâce à des dictionnaires de valences, d'associer chacun des mots avec un sentiment, et d'y adjoindre d'autres caractéristiques (telles que le nombre de mots de négation ou encore le nombre de mots concernant la cible suivis de mots positifs) comme donnée d'entrée pour un classifieur automatique SVM. Ce système hybride, exploitant un dictionnaire de valences pour calculer des caractéristiques d'entraînement sur un corpus, atteint une fiabilité de 89% sur un corpus de 940 tweets en anglais. Le résultat est ensuite exploité par une technique de résumé de sentiment : parmi le corpus de messages, une sélection est proposée, qui propose les informations les plus pertinentes au lecteur.

[Khan et al., 2015] a proposé un algorithme en plusieurs étapes, illustré en figure 2.8 : les messages, des tweets, sont triés entre trois types de phrases : déclaratifs, impératifs, interrogatifs. Ensuite, des outils linguistiques résolvent les coréférences, agrégent les opinions, et enfin entrainent un classifieur. Leurs étapes de prétraitement sont assez lourdes (correction complète des orthographes inventives, élimination des URL, hashtags et mentions, ...). Ils atteignent une $F_1 = 0.749$ en moyenne sur 5 thèmes, constituant un corpus de plusieurs centaines de milliers de tweets.

8. *Recursive Neural Network*, une architecture d'apprentissage profond.
9. `https://nlp.stanford.edu/sentiment/index.html`

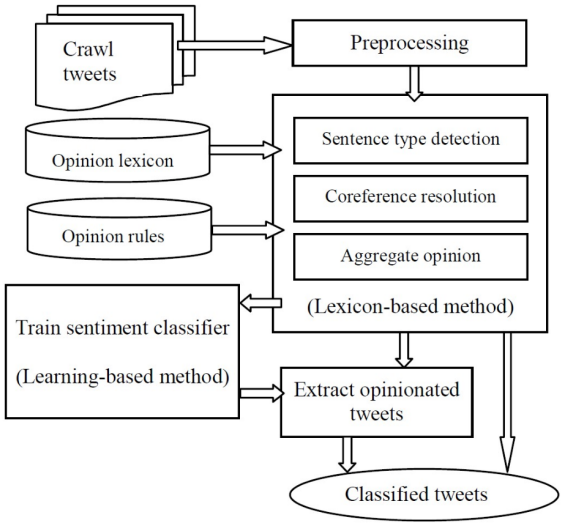



FIGURE 2.8 – Système hybride proposé par [Khan et al., 2015] (Fig.1)

## 2.5 Détection de l'opinion, de la posture

La détection d'opinion se focalise sur le couple (cible, polarité) qui convient au mieux à un texte. La polarité est alors relative à la cible, ce qui diffère parfois du sentiment général du texte. Pour procéder à la détection, deux approches existent. La première, d'*opinion mining*, consiste à repérer le sujet du texte, inconnu *a priori*. La seconde, de *stance detection*, part d'une cible *a priori*, et cherche à déterminer si le message y est favorable ou non.

### 2.5.1 Détection de la cible et de l'expression de l'opinion

La détection ou extraction de l'opinion peut être vue et réalisée grâce à un système composé de deux modules principaux, illustrés en figure 2.9 : d'une part, le sujet d'un texte est extrait, par exemple avec des techniques issues du domaine de l'Extraction d'Information. Ce sujet devient la **cible** de l'opinion. En parallèle, parfois en dépendance, le sentiment du texte est calculé, et devient la **polarité** de l'opinion.

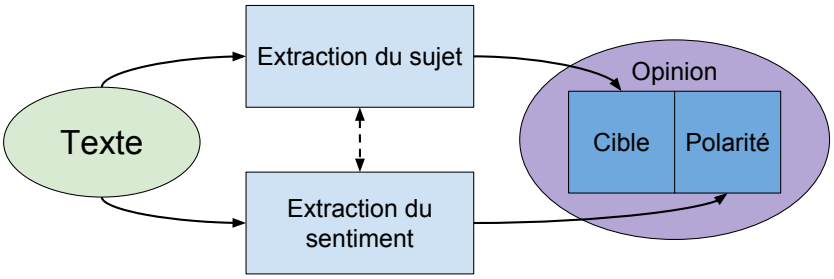



FIGURE 2.9 – Extraction classique des opinions d'un document

[Liu et al., 2015] a proposé d'utiliser un réseau neuronal, avec en entrée des « word embeddings » ; le réseau attribue des labels aux mots dans la phrase : « TARG » pour la cible, et labels d'expression de sentiment « EXPR », sur les mots. Un exemple est donné en table 2.3. La première ligne montre le texte analysé ; la seconde montre la sortie prédisant la cible ; la troisième ligne signale la présence d'expressions de la polarité. Dans cet article, plusieurs approches de réseaux de neurones profonds

sont comparées : réseaux récurrents (RNN), *conditional random fields* (CRF) et *bidirectionnal long-short term memory-RNN* (Bi-LSTM-RNN). Ce dernier atteint une F1 de 0,80 sur un corpus en anglais, composé de revues de clients à propos d'ordinateurs et de restaurants.

TABLEAU 2.3 – Un exemple tiré de [Liu et al., 2015], annoté pour la cible (TARG) et pour l'expression de la polarité (EXPR)

| The | hard | disk | is | very | noisy |
|---|---|---|---|---|---|
| O | B-TARG | I-TARG | O | O | O |
| O | O | O | O | B-EXPR | B-EXPR |

Une particularité de ce type de réseau (LSTM) repose dans la manière dont la donnée est fournie au classifieur. Alors qu'un algorithme d'apprentissage classique, par exemple un SVM, ingère des vecteurs de taille fixe, les réseaux testés ici parcourent le texte d'entrée mot à mot, avec une fenêtre de taille variable (une taille de 3 semble optimale selon [Liu et al., 2015]). Ainsi, le classifieur prédit un label pour chacun des mots, permettant d'identifier la présence de la cible ou de la polarité terme à terme. Notons que ce réseau ne prédit pas *quelle* est la polarité du texte.

Dans la définition d'une opinion de [Pang and Lee, 2008], l'un des éléments est souvent sous-exploité : il s'agit de l'*aspect* de la cible. Des travaux visent à détecter l'aspect sur lequel porte le sentiment, dans des revues de produits [Poria et al., 2016]. Ils classifient les mots en « aspect » ou « non-aspect » uniquement, et revendiquent d'excellents scores (85% en iso-précision-rappel). Comme précédemment, l'approche suivie est basée sur l'apprentissage profond (ici à base de réseaux convolutifs, CNN), couplé à des patrons linguistiques : un ensemble de 5 règles pré-définies, considérant la valeur grammaticale des mots.

Dans ces deux travaux, les réseaux de neurones déployés ont nécessité des millions de documents annotés dans les corpus d'apprentissage, notamment des revues de produits sur Amazon. Ils sont alors dédiés à l'analyse de sentiment dans des revues de produits, qui constitue un format d'expression écrite assez particulier. Enfin, le recours à des règles linguistiques suggère qu'il peut être plus efficace d'exploiter tant les ressources linguistiques disponibles, que la puissance de l'apprentissage. Ces règles ont souvent un rappel plus faible, mais qui est compensé par une meilleure précision que l'apprentissage « pur ».

### 2.5.2 Classification d'opinion : au-delà de la cible, la posture

L'infinité des possibilités d'attribution de la cible lors de son extraction rend complexe tant la validation des méthodes précédentes, que leur exploitation : pour obtenir une vue d'ensemble sur la réputation d'une marque, par exemple, il faut réussir à agréger intelligemment les opinions exprimées sur tous les aspects des produits, personnes, organisations soutenant ou rivalisant envers la marque-cible. Ainsi, il est parfois plus intéressant de fixer la cible *a priori*, d'en relâcher le sens, et d'analyser les textes en regard de cette cible. Au-delà des enquêtes de satisfaction-client à vocation commerciale, l'analyse de textes porteurs d'opinions politiques vise souvent à attribuer un « camp » aux auteurs de messages. Ceux-ci ne sont plus vraiment analysés en regard d'une cible, mais envers le camp, synthétisé par la cible. Quelques travaux entièrement tournés vers la détection d'opinions plus politiques, ou provenant de débats, apportent un éclairage nouveau à ce problème avec la notion de posture.

[Tsytsarau and Palpanas, 2012] propose, au sein d'un article d'état de l'art, de transformer la tâche de détection de posture en détection de contradiction. En utilisant des messages provenant d'un forum, il constate qu'il y a souvent des signes d'opposition ou d'agrément entre deux publications successives d'un même fil de discussion, sans nécessairement de marqueur explicite du camp soutenu ou dénigré. Un constat similaire résulte de travaux sur des débats en ligne [Anand et al., 2011], qui distingue les messages en termes d'opposition à des entités : la majorité des posts sont de sentiment négatif, et dénigrent le parti opposé. Il n'y a parfois pas de référence absolue dans le texte : la cible n'est pas

explicitement mentionnée, mais est relative au message précédent. La polarité ou la posture est alors difficile à comprendre sans prendre en compte les éléments antérieurs dans le fil de discussion.

Une étude portant sur des textes politiques britanniques [Maynard and Funk, 2011] utilise GATE [Cunningham et al., 2011], un moteur de règles linguistiques, pour identifier des opinions politiques. Un ensemble de règles détermine l'adhésion ou non à l'un des trois principaux partis politiques au Royaume-Uni.

[Andreevskaia and Bergler, 2008] a combiné les deux approches les plus courantes : utilisant des dictionnaires de sentiment, ainsi que de l'apprentissage automatique. Dans son expérience, il propose un corpus de messages issus d'un forum de débats (par exemple, « Firefox VS Internet Explorer »). Les messages sont ici considérés comme des éléments d'une discussion, ce qui est pris en compte par leur classifieur.

[Hasan and Ng, 2013] atteint une fiabilité de 75% sur une tâche de détection de posture sur des forums politiques (fiabilité variant de 70,9% à 75,4% selon la thématique). Plusieurs approches sont explorées : la présence d'unigrammes et bigrammes de mots, des caractéristiques provenant de la ressource LIWC, ainsi que des caractéristiques spécifiques à l'instance de la tâche : une « contrainte idéologique » force les utilisateurs à conserver une même posture tout au long d'un débat ; une « contrainte d'interaction » suppose que deux messages successifs ont une plus grande probabilité d'être opposés. Notons que la construction de ces caractéristiques découle d'hypothèses assez fortes, qui peuvent fonctionner sur une instance sans pour autant être généralisables.

TABLEAU 2.4 – Comparatif des approches de détection d'opinions politiques

| Référence | Corpus | Approche | Score |
|---|---|---|---|
| [Andreevskaia and Bergler, 2008] | revues de films, presse | hybride | Acc=71,1% |
| [Anand et al., 2011] | débats, convinceme.net | classifieur | Acc=66,9% |
| [Hasan and Ng, 2013] | débats, convinceme.net | hybride | Acc=75,4% |
| [Maynard and Funk, 2011] | tweets politisés | règles | P=62%, R=37% |

Les scores des différents travaux mentionnés ici sont montrés en table 2.4. Nous remarquons des niveaux de performance très nettement inférieurs à ceux atteints sur des revues de produits par des analyseurs de sentiment plus « classiques », et ce malgré des techniques parfois très adaptées aux corpus exploités. La détection de posture présente donc un défi particulier, distinct bien que proche de l'analyse du sentiment.

## 2.6 Synthèse sur l'analyse de sentiment et l'extraction d'opinion

Afin d'extraire de l'information quant à l'appréciation des utilisateurs envers des produits, films, marques ou entités politiques, les quatre notions, émotion, sentiment, opinion et posture, sont utiles. La volonté de résumer des textes est justifiée et pertinente, au regard de la quantité de messages émis chaque jour, et de l'impact que certains peuvent avoir. La détection de l'*opinion* répond à ce besoin, et permet une analyse fine tant sur la répartition temporelle, par auteur, que par cible, aspect de l'opinion ; sur une cible donnée, l'évolution des polarités permet d'obtenir des indicateurs sur la popularité d'un produit, d'un candidat politique, d'un film.

L'identification de l'émotion, plus fine, permet de mieux caractériser la polarité de l'opinion : un utilisateur malheureux est-il triste, ou en colère ? Cette finesse se fait au détriment de la performance des modules d'analyse, rendue difficile par le grand nombre de classes, avec souvent huit émotions. Réduisant ce nombre de classes, le sentiment, parfois trop générique, permet d'attribuer un score à tous les textes et donc de bénéficier de statistiques à l'échelle de la collecte des messages.

Pour pallier cette généricité, un classifieur plus précis sur un domaine peut être entraîné pour déterminer la posture d'un texte, en structurant plus fortement le problème, et résultant en des performances plus fiables sur le domaine en question ; cependant il est rare d'atteindre des F1-mesures

supérieures à 0.70. Cette tâche de détection de posture retient notre attention, car elle peut s'appliquer sur un grand nombre de sujets d'intérêt où le sentiment donne des résultats trop mitigés.

TABLEAU 2.5 – Synthèse des ressources et approches pour les différentes notions

| | Dictionnaires | Règles | Apprentissage & Hybride |
|---|---|---|---|
| Émotion | [Mohammad and Turney, 2010]<br>[Bradley and Lang, 1999] | [Munezero et al., 2015] | [Yassine and Hajj, 2010] |
| Sentiment | [Baccianella et al., 2010]<br>[Strapparava et al., 2004] | [Khan et al., 2014]<br>[Hutto and Gilbert, 2014] | [Pang et al., 2002]<br>[Pak and Paroubek, 2010] |
| Opinion | [Wiebe et al., 2005] | [Liu et al., 2015]<br>[Poria et al., 2016] | [Pang and Lee, 2008] |
| Posture | | [Maynard and Funk, 2011]<br>[Tsytsarau and Palpanas, 2012] | [Hasan and Ng, 2013] |

Chacune des quatre notions est propice à l'élaboration d'approches novatrices ; cependant nous avons identifié deux grandes familles d'étiqueteurs. La première famille, des analyseurs basés sur des dictionnaires, attribue des valeurs intrinsèques à chaque mot ; une combinaison de ces valeurs produit la polarité recherchée. La seconde approche nécessite la constitution d'un corpus d'entraînement à partir duquel sont extraites des caractéristiques, par exemple des sacs de mots. Cela permet le calcul d'un modèle de prédiction, dont le SVM et Bayes naïf sont les implémentations favorites. Nous agrégeons l'approche par apprentissage, et les approches hybrides. Cette vue du domaine est présentée en table 2.5, et replace les travaux principaux mentionnés dans ce chapitre.

La difficulté de l'analyse est intrinsèque au texte : quelques mots permettent d'exprimer de nombreuses idées, parfois opposées. Dans notre cas, les tweets sont très courts, et l'implicite y a une grande place. De plus, l'utilisation de la ponctuation, d'émoticônes, de hashtags, de photos et d'URLs, parfois ayant un rôle sémantique, augmente la difficulté. Enfin, les argots (parfois nommés sociolectes, dialecte de réseau social) sont très présents, nous privant souvent des ressources linguistiques nécessaires. Les mots nouveaux, en usage seulement sur Twitter, sont souvent considérés comme des labels sans prendre en compte la relation avec le contexte, sans autoriser un sens multiple.

Dans le chapitre 5, nous nous pencherons sur la tâche de détection de posture, à partir de tweets. La posture est souvent plus exploitable que le sentiment, mais aussi plus difficile à extraire ; de plus, peu de travaux portent sur sa détection dans des tweets.

CHAPITRE 3

# Étude des comptes utilisateurs

L'étude de l'espace médiatique nécessite de caractériser les auteurs, en sus de l'analyse des messages émis. Le second axe de notre recherche procède ainsi à l'analyse des comptes-utilisateurs, afin de répondre au « qui ? » de Lasswell [Lasswell, 1948]. Sur l'espace médiatique qui nous intéresse, les contenus semblent parfois issus d'une foule de comptes anonymes ; or ce n'est pas « les réseaux sociaux » qui discutent et débattent, mais bien des comptes identifiables et caractérisables. Pour éclaircir cette complexité, deux grandes approches s'offrent à nous.

La première approche consiste en l'établissement d'un **profil** d'utilisateur : il faut rassembler l'ensemble des données concernant un compte utilisateur, afin d'en caractériser son comportement. Dans la littérature, de tels profils, bien que fréquents, différent d'une application à l'autre. Cette structure de données sert ensuite à comparer les actions réalisées par deux comptes différents ou à élaborer des modèles de classification. Nous abordons cette analyse du comportement en section 3.1.

Une deuxième approche attribue un score d'importance ou d'**influence** aux comptes. L'influence est une notion floue, dont la définition évolue dans la littérature. Dans la section 3.2, nous présentons des travaux considérant l'influence comme l'importance dans la propagation d'information ; à suivre dans la section 3.3, l'influenceur est celui en qui est placée la confiance des membres du réseau. Enfin, nous explorons les notions de réputation et de position sociale en section 3.4 : le réseau social numérique est vu comme un graphe, dont les nœuds sont les comptes utilisateurs. Sur ce graphe est mesurée l'importance d'un nœud par l'évaluation de ses caractéristiques topologiques ; ces mêmes caractéristiques alimentent un modèle d'apprentissage non-supervisée, déterminant la position-type, ou rôle du nœud dans le réseau.

Pour ces deux approches, nous présentons l'intuition derrière les modèles et les modèles eux-mêmes, ainsi que le détail des mesures ou classifications obtenues, couvrant l'ensemble des types de données disponibles aidant à caractériser les comptes, auteurs médiatiques. Enfin, une synthèse conclut ce chapitre.

## 3.1 Construction d'un profil de comportement

La caractérisation des comptes nécessite des données, souvent constituées des actions accomplies par les utilisateurs. Nous voyons dans cette section quelles données sont collectées et analysées, et comment se mesurent et se comparent les comportements des comptes.

Nous passons en revue différentes applications nécessitant la construction d'un format de données spécifique pour représenter les comptes utilisateurs ; dans la littérature, cette structure de données porte le nom de « profil » ; sa construction est parfois appelée *cyber-profilage*.

### 3.1.1 Profil psychologique

Une analyse du comportement d'utilisateurs de Facebook et de Twitter a été menée via le crowdsourcing, sur la plate-forme AMT [1] [Panek et al., 2013]. L'objectif visé, d'analyse psychologique, consiste en la quantification du comportement narcissique, comparée entre les deux plate-formes. Les données recueillies, concernant une centaine de personnes, ont ensuite servi à l'extraction de six « clusters psychologiques » liés à des types de comportement narcissiques (par exemple, l'exhibitionnisme, la supériorité, l'auto-suffisance). Les caractéristiques recueillies se composent de l'activité (nombre de connexions par jour, durée des sessions de lecture, quantité de messages postés) et des réponses à un questionnaire psychologique, le *Narcissism Personality Inventory* NPI-16.

L'intérêt d'une analyse psychologique (parfois décrite sous le nom *psychographics*) dépend du modèle retenu : si le narcissisme ne correspond pas à l'application choisie, un modèle plus général, *OCEAN* (initialement appelé *Revised NEO Personnality Inventory*), décrit la personnalité des individus [Costa Jr et al., 1991]. Ce modèle distingue les traits de caractères (confiance, altruisme, modestie...) des aspects de facettes plus englobantes, au nombre de cinq :

1. **O** : *Openness*, ouverture d'esprit : mesure la propension de la personne à la curiosité, si elle est avide de nouveautés, d'expérimenter ;
2. **C** : *Conscientiousness*, conscience : mesure la réflexion de la personne sur ses actions, pour la positionner entre réactif (spontané) et délibératif (organisé, planificateur) ;
3. **E** : *Extraversion*, extraversion : évalue la facilité de la personne à aller vers les autres, si elle est solitaire ou sociale. Il s'agit du volume d'actions interpersonnelles désirées par la personne ;
4. **A** : *Agreeableness*, agréabilité : quantifie la confiance et l'altruisme dont est capable la personne, elle correspond à la qualité des actions interpersonnelles ;
5. **N** : *Neuroticism*, neuroticisme : mesure la stabilité émotionnelle de la personne, ainsi que son impulsivité.

Le questionnaire, en 240 items, permet d'attribuer des valeurs numériques à chacune des cinq facettes. Sur une population de 394 personnes, les corrélations entre chacune des différentes facettes sont faibles ; seules les paires extraversion-ouverture, et neuroticisme-conscience présentent un lien ($|r| > 0.3$).

Le modèle *OCEAN* a été exploité pour analyser le lien entre personnalité et usage des réseaux sociaux numériques [Quercia et al., 2011]. *myPersonality*, une application Facebook, a collecté les traits de caractère de nombreux utilisateurs volontaires, via un questionnaire. Parmi ces volontaires, 355 ont associé leurs comptes Facebook et Twitter : les auteurs comparent alors les caractéristiques O, C, E, A et N, avec les statistiques des comptes Twitter (nombre de followers, d'abonnements, de listes), et montrent qu'une régression sur ces statistiques d'utilisation de Twitter prédit le profil psychologique de manière fiable.

### 3.1.2 Profil des thématiques abordées par un compte

Motivé par l'amélioration des systèmes de recommandation pour aider les utilisateurs à trouver de meilleurs contenus ou contacts, le service TUMS (*Twitter-based User Modeling Service*) propose de représenter l'usage fait par les comptes des « concepts », éléments de C l'ensemble des « concepts d'intérêt », c'est-à-dire les hashtags, entités nommées et thématiques émises [Tao et al., 2011]. Dans ces travaux, un profil d'utilisateur est défini comme un ensemble de poids liés à chacun des « concepts ». Les schémas de pondération recommandés sont soit « *term-frequency* », soit « *term-frequency, inverse document frequency* » : le premier compte le nombre d'occurrences du concept dans les messages émis par l'utilisateur ; le deuxième divise ce nombre par le nombre total d'occurrences du concept dans le corpus.

1. Amazon Mechanical Turk : les utilisateurs étaient rémunérés pour être monitorés.

Ces comptages sont appliqués à trois types d'items différents : hashtags, entités nommées et thématiques. Ainsi, les auteurs représentent les centres d'intérêt d'un compte par un « tableau de bord » constitué de graphiques illustrant la distribution de l'usage des items. Cette représentation des comptes est considérée trop limitée par son approche thématique seule, aussi un modèle plus inclusif, qui ajoute la dynamique temporelle, a été introduit [Yin et al., 2014]. Nous y reviendrons en section 3.1.5.

### 3.1.3 Détection de robots

De larges pans de l'activité légitime sur Twitter est constitué de robots, souvent mus par des entreprises ou institutions, répliquant de manière automatisée sur plusieurs réseaux sociaux un même contenu. Leur proportion pourrait atteindre 15% des inscrits sur le réseau en 2017 [2] soit 48 millions de comptes ; au minimum elle serait d'au moins 8,5%, suivant différentes motivations : *spambots*, *paybots* et *influence bots* [Subrahmanian et al., 2016]. Ce découpage n'est pas exhaustif, mais il correspond aux principaux objectifs malicieux recherchés :

— *spambots :* la définition la plus large, elle englobe tous les diffuseurs de messages non désirés, qu'ils soient liés à des escroqueries ou non ;

— *paybots :* ces robots diffusent des contenus légitimes, mais via des adresses redirigées à travers des sites rémunérant le trafic généré. Ainsi le créateur du robot est rémunéré au nombre d'utilisateurs suivant son lien, vers un contenu qui ne lui appartient pas ;

— *influence bots :* la vocation de ces comptes est de propager une information, un point de vue, vis-à-vis d'un problème donné ; les applications incluent des choix sociétaux (par exemple : pour ou contre la vaccination, l'avortement), électoraux (comme la promotion ou le dénigrement d'une personnalité politique, d'un parti), ou publicitaires (tel que la promotion ou le dénigrement d'une entreprise, d'une industrie).

Les robots ou faux comptes sont aussi appelés comptes *sybils*, ainsi nommés à cause de la dissociation entre les identités revendiquée et réelle. Afin de les identifier automatiquement sur le réseau social chinois Renren, un classifieur SVM entraîné sur 1 000 « sybils » et 1 000 « vrais comptes » atteint une fiabilité de 99% [Yang et al., 2014]. Les données utilisées excluent le texte, se focalisant sur l'usage des fonctionnalités de la plate-forme : taux d'acceptation des requêtes d'amis, séparément selon qu'elles soient émises ou reçues, fréquence d'envoi de requêtes d'amitié (en quantité par heure), et coefficient de clustering dans le graphe constitué par les 50 premiers amis d'un compte. Ce système est tout à fait adapté pour détecter des comportements de spammeurs agressifs, reposant sur l'ajout massif de connexions (amitié).

Une tâche de détection de robots plus subtils, intitulée « *DARPA bot challenge* », a encouragé des progrès dans la catégorisation des comptes, favorisant l'émergence de bonnes pratiques et d'une liste de caractéristiques d'intérêt, exploitables pour entraîner un classifieur « bot » *versus* « compte légitime » [Subrahmanian et al., 2016]. Sur un corpus artificiel de 7 038 comptes au format Twitter, les participants devaient trouver les *influence bots*, dont le nombre (40) leur était inconnu. Les caractéristiques choisies par les différents concurrents recouvrent les aspects suivants :

— la syntaxe : nombre de hashtags, d'URLs, de caractères spéciaux ; proportion de messages finissant par un hashtag ou une URL ;

— la sémantique : proportion de messages portant sur un même sujet, sentiment moyen, nombre de langues utilisées ;

— le comportement temporel : évolution du sentiment dans le temps, analyse des écarts temporels entre deux émissions de tweets, nombre moyen de tweets par jour, évolution du nombre d'abonnés ;

2. https ://www.cnbc.com/2017/03/10/nearly-48-million-twitter-accounts-could-be-bots-says-study.html

— le profil du compte : taux de remplissage du profil, apparence du pseudonyme (présence de tiret bas, réplication de patrons génératifs de pseudonymes), nombre de messages, mentions, etc. ;

— la topologie du réseau : degré du compte dans les graphes d'abonnements, de mentions ; ressemblance d'un sous-réseau avec un réseau aléatoire, suggérant un groupe de robots.

Au cours du challenge, d'une durée de quatre semaines, les modèles mis au point par les équipes ont évolué. Le nombre de caractéristiques différentes prises en compte par l'équipe Sentimetrix, par exemple, est passé de 66, pour le modèle initial, à 175 en fin de défi ; ces informations étaient fournies à un classifieur SVM pour prédire le type de compte : *bot, or not*.

La meilleure fiabilité obtenue lors du « *DARPA bot challenge* » résulte de l'exploitation humaine d'une détection d'observations aberrantes (*outliers*). La méthode complète est la suivante : le jeu de données, où chaque individu est représenté par les caractéristiques retenues, subit tout d'abord une phase de réduction de dimension (méthode NMF, Factorisation Non-négative de Matrice), puis alimente un algorithme de *clustering* qui distingue nettement la foule d'utilisateurs « normaux », des anomalies recherchées. Ainsi se trouve réduite la longueur de la liste des comptes sur lesquels les enquêteurs humains travaillent, vérifiant l'humanité des uns, bannissant les autres.

### 3.1.4 Distances entre comptes utilisateurs

Une liste de caractéristiques comparable a été élaborée pour une autre fin : l'identification d'un même individu derrière plusieurs comptes, répartis sur plusieurs plate-formes de médias sociaux [Liu et al., 2014]. La méthode n'adresse pas la présence multiple d'un individu sur une même plate-forme. La figure 3.1 illustre leur méthode : une collecte d'informations provenant de plusieurs réseaux sociaux différents, occidentaux (Twitter et Facebook) et chinois (Renren et Sina Weibo) pour les principaux, vient enrichir des profils multi-aspects. Des modules de comparaison spécifiques (par exemple, pour comparer deux images, qui peuvent être la même photo à différentes résolutions ou cadrages) alimentent un score de similarité entre profils, permettant éventuellement de les lier s'ils correspondent à la même identité réelle.

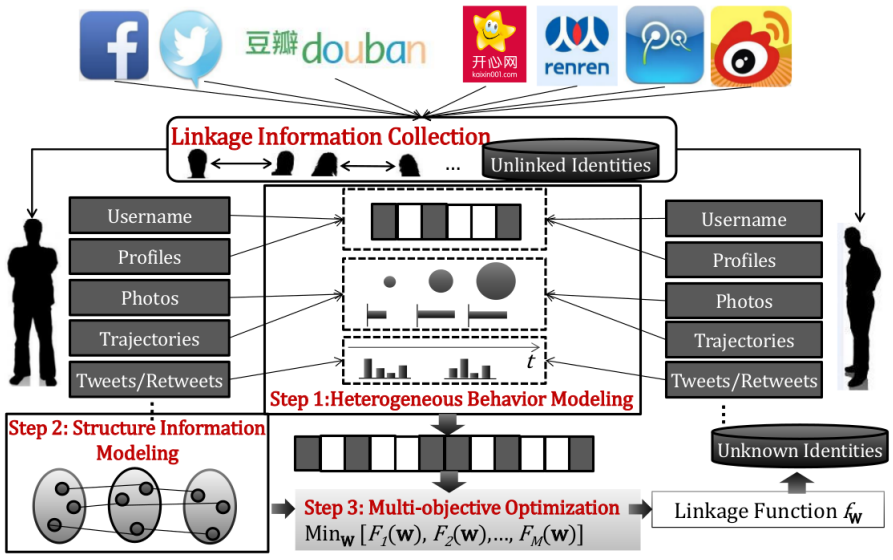



FIGURE 3.1 – Comparaison de profils issus de plate-formes différentes, selon [Liu et al., 2014]

Ce système porte plus sur le stockage et l'élaboration et adaptation de comparateurs pour chacune des modalités (profils, photos, textes, trajectoires) à travers des réseaux sociaux différents, dont les schémas de données divergent. Cette adaptation n'a rien de trivial : les images, textes et informations diverses ne sont pas structurées de la même façon et ne sont pas forcément présents de façon

équilibrée sur les plate-formes numériques observées. Enfin, la pondération des scores des différents comparateurs permet d'atteindre le but recherché : par exemple, la comparaison des images suggère une proximité entre deux comptes ; ce signal est renforcé par les similarités dans les informations de profils, et permet au classifieur de statuer.

Certaines pages de l'encyclopédie Wikipedia sont sujettes à des modifications malicieuses, ne reflétant pas la réalité. Afin de faciliter la tâche des modérateurs du site, un modèle de classification détermine si un compte est *sybil* ou non. Cependant, les modes d'attaque actuels reposent sur des identités multiples qui effectuent de manière anodine les changements. Visant à regrouper les faux comptes manipulés par un même individu (dits *sockpuppets*, poupées de chiffon), le système *SocksCatch* construit un graphe bipartite (éditeurs, pages) ; une distance est proposée entre les nœuds, prenant en compte la similarité des actions d'édition des pages, et leur proximité dans le temps [Yamak et al., 2018].

Dans ces travaux, nous avons vu comment se structurent les modèles représentant globalement les comptes utilisateurs ; nous nous penchons plus en détail dans le paragraphe suivant sur la gestion des informations temporelles.

### 3.1.5 Travaux sur la dynamique du comportement

La dynamique du comportement apporte des informations déterminantes pour caractériser le comportement. Elle peut être représentée comme l'histogramme de l'activité du compte, décrivant la quantité d'actions et modifications accomplies heure par heure sur une journée, ou sur une semaine [Liu et al., 2010]. Ainsi, sur le site Google Actualités, le nombre de clics par mois pour chacune des rubriques est compté, d'où s'obtient un vecteur de fréquence dont les dimensions sont les rubriques. Ce vecteur représente, pour un mois donné, les préférences d'un utilisateur et permet d'en visualiser l'évolution sur une année, et de les agréger au niveau national par exemple. Les auteurs constatent des écarts significatifs sur les distances entre profils selon le pays affiché : les habitudes de navigation de la population de chaque pays ne sont pas homogènes, mais diffèrent nettement de celles du pays voisin.

Un tel modèle, combinant thématique et temporalité, auxquelles s'ajoute la dimension spatiale (combinant donc contenus, date d'émission, lieu d'émission), a été proposé pour tirer parti des réseaux sociaux géographiques tels que FourSquare et Loopt (*location-based social networks*). Sur ces plate-formes, les utilisateurs signalent leurs entrées dans des lieux d'intérêt : musées, parcs, restaurants, bureaux, plus rarement le domicile. Les auteurs développent un système de recommandation des lieux d'intérêt, correspondant au mieux aux préférences thématiques de l'utilisateur, mais aussi au moment de la requête dans la journée, dans la semaine, ainsi qu'à l'éloignement du domicile (ce qui différencie les vacances de la routine) [Yin et al., 2016].

Plus la granularité est fine, et plus le stockage nécessaire est important, ce qui devient problématique lorsque l'analyse porte sur une grande population. Plutôt que de se limiter aux heures exactes, il peut être pertinent de ne diviser la journée qu'en 4 périodes, ce qui suffit à prédire l'activité d'un compte et détecter d'éventuelles anomalies. Cette affirmation provient d'une analyse de Twitter où quatre types d'actions sont répertoriées :

— *posting* : émission d'un message, *tweet* ;
— *forwarding* : transfert d'un message, c'est-à-dire *retweet* ;
— *liking* : marquage d'une publication comme « aimée », anciennement « favorite » ;
— *replying* : publication d'une réponse envers un message original posté par un tiers.

Le modèle ne compte que les actions de *posting*, dont le sujet (présence de mots-clés définis *a priori*) et le sentiment sont extraits automatiquement. La constitution de tels profils, par l'étude

d'une population (les comptes interagissant avec une personnalité politique), fournit des probabilités d'enchaînement d'actions vraisemblables pour un système multi-agents simulant Twitter [Gatti et al., 2013].

Afin de prédire les périodes d'activité des comptes, une modélisation par des chaînes de Markov cachées utilise les observations (actions émises par le compte, actions reçues) pour déterminer l'état : actif (connecté) ou non [Raghavan et al., 2014]. Un jeu de données réelles permet d'apprendre des paramètres, plus précisément des ratios :

- $\frac{p_1}{p_0}$ représente la propension d'un compte à réagir aux mentions le concernant (qui incluent les retweets et les réponses) :
  - $p_1$ est la probabilité que l'utilisateur s'active, sachant qu'il a été mentionné dans l'intervalle de temps précédent ;
  - $p_0$ est la probabilité que l'utilisateur ne s'active pas, sachant qu'il a été mentionné dans l'intervalle de temps précédent ;
- $\frac{\gamma_1}{\gamma_0}$ correspond à la capacité d'un compte à rendre actif ses voisins :
  - $\gamma_1$ est la probabilité que l'entourage d'un utilisateur s'active, sachant que lui-même est actif ;
  - $\gamma_0$ est la probabilité que l'entourage d'un utilisateur s'active, sachant que lui-même n'est pas actif.

La figure 3.2, issue de [Raghavan et al., 2014], illustre la distribution de comptes réels. Les lignes en vert pointillé ne présagent pas du modèle utilisé, elles ne servent qu'à clarifier la séparation entre les clusters. Les auteurs en déduisent un découpage en trois catégories de comptes :

1. le cluster 1, autour de (1 ;1), en noir : il représente une majorité de comptes, qui reçoivent et réémettent sans déséquilibre :
2. le cluster 2, en bleu, vers la droite de la figure : ces comptes s'activent lorsqu'ils sont mentionnés par leurs contacts ;
3. le cluster 3, en rouge : ces comptes déclenchent l'activité de leurs abonnés.

La présence d'une tendance temporelle cyclique, à prendre en compte dans la représentation de l'activité d'un compte au long d'une journée est attestée [Yuan et al., 2013], résultant en des patrons, par exemple des profils moyens d'activité. Les données, issues de réseaux sociaux géographiques (chaque contenu publié est associé à des coordonnées GPS), sont représentées sous la forme d'un cube « UTP », pour Utilisateur - Temps - Point d'intérêt. Chaque élément $c_{u,t,p}$ du cube vaut 1 si l'utilisateur $u$ est présent au point $p$ durant l'heure $t$.

Ce cube est alors stocké sous la forme d'une matrice creuse ; cependant il permet aussi de récupérer rapidement les fréquentations usuelles selon chacun des axes. À une heure donnée, où se trouvent les utilisateurs ? Quelle est la quantité de personnes présentes dans le lieu $p$ au fil de la journée ? Comment $u$ organise-t-il sa journée ? De plus, une fonction est proposée pour calculer la similarité de deux lieux, et de deux comptes, ce qui aide à améliorer un système de recommandation : il s'agit de la similarité de cosinus (*cosine similarity*), qui compte la proportion de lieux identiques visités par deux comptes, par rapport au produit du nombre de lieux visités par chacun des deux comptes. Pour ce faire, les lieux sont vus comme des mots, l'ensemble de l'activité d'un utilisateur constituant un document : la dimension P ressemble alors à un vecteur de fréquences de termes. La comparaison entre deux utilisateurs nécessite un vocabulaire commun, aussi les comparaisons se font à l'intérieur d'une même ville.

Cette similarité permet de positionner les signaux personnels les uns en fonction des autres, et ainsi d'en extraire des groupes de comportement similaires, qu'il s'agisse de groupes d'individus ou de groupes de lieux.

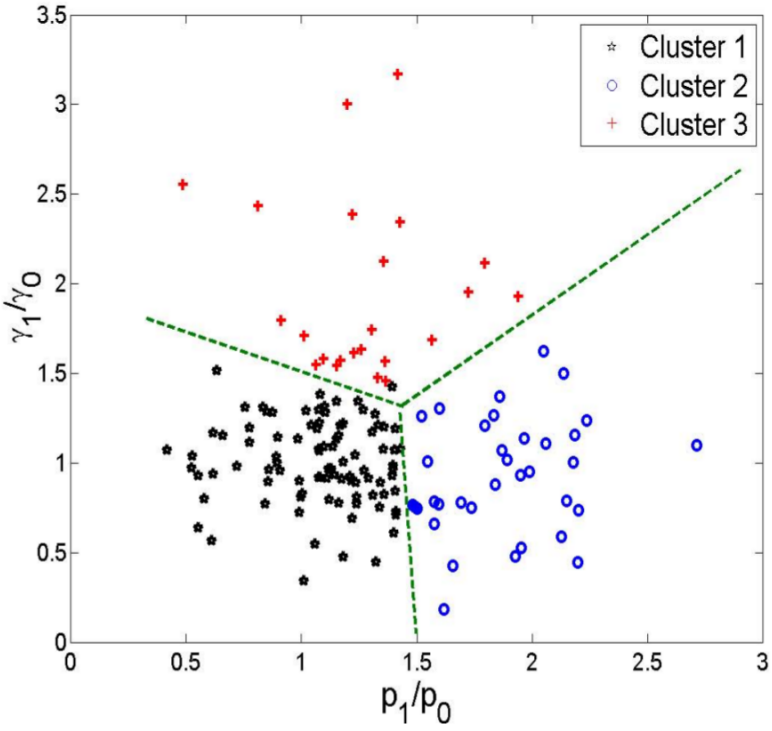



FIGURE 3.2 – Catégorisation non supervisée de comptes selon leurs réactions temporelles [Raghavan et al., 2014]

### 3.1.6 Discussion sur la construction de profils-utilisateurs

La revue de ces travaux clarifie les diverses implémentations de la notion de *profil* : il s'agit d'une structure de données, associée à une distance, et en conséquence à un calcul de similarité. Nous avons parcouru les grandes approches de mesure des actions individuelles, débouchant sur la constitution de profils comportementaux. Les caractéristiques retenues dépendent de l'application visée ; notons que pour un même problème (par exemple, la détection de robots), ces listes évoluent. D'une part, parce que les plate-formes évoluent, ouvrant et fermant de nouvelles fonctionnalités et contrôles ; d'autre part, parce que l'application évolue (les robots de 2018 ne sont pas les mêmes que ceux de 2008).

Dans la section suivante, nous abordons la notion d'influence : plutôt que de regrouper toutes les informations disponibles sur un compte, il s'agit « uniquement » d'évaluer son importance. En effet, sur un réseau social numérique, chacun a droit à la parole. Pourtant, tous n'ont pas le même impact, et certains comptes occupent une place centrale dans le débat : ils sont appelés les *acteurs-clés*. En section 3.2, nous commençons par définir l'influence, puis nous nous penchons sur un premier type d'acteurs-clés : ceux dont le rôle est prépondérant dans la diffusion de l'information.

## 3.2 Influence dans un modèle de diffusion de l'information

Le dynamisme des échanges et la quantité d'émetteurs de messages dans les réseaux sociaux numériques rend nécessaire l'utilisation d'outils pour déterminer les acteurs influents, parfois appelés *acteurs-clés* : il s'agit de ceux qui initient la propagation d'une information ou d'une opinion, qui occupent un rôle crucial dans leurs (non-)propagations. La recherche d'acteurs-clés dans un réseau social nous mène naturellement à la notion d'*influence*, qui couvre plusieurs facettes en définition 3.2.1.

**Definition 3.2.1.** *Influence :* potentiel, détenu par une entité, de faire réaliser par une autre entité une action, ou de lui faire atteindre un état. Cette action ou cet état peuvent être désirés. Nous clarifions

les deux résultats :

— réaliser une action. L'influenceur détient alors une capacité à faire réagir les autres à son discours ou attitude.

— faire atteindre un état. Il s'agit, par exemple, de l'état « informé », d'un état sentimental ou émotionnel, ou autre. L'influenceur est alors un émetteur ou un relais principal de l'information à laquelle l'influencé réagit.

L'influence est ainsi liée à la diffusion de l'information dans un réseau social, qui est un phénomène complexe : tant les sensibilités des gens que l'intensité de leurs relations entrent en jeu.

Les diffusions d'information ou d'opinion sont différentes : une phrase suffit à informer quelqu'un, si la phrase est lue ; la propagation d'opinion est un phénomène plus complexe, qui consiste à convaincre ou persuader l'autre. Dans cette section, nous nous limitons à l'information, pour laquelle plusieurs modèles de propagation existent. Nous les présentons ci-dessous.

### 3.2.1 Modèles de diffusion d'information dans les réseaux sociaux

Les modélisations de la diffusion de l'information dans un graphe représentant un réseau social se basent principalement sur deux approches : *Independent Cascades* [Goldenberg et al., 2001] et *Linear Threshold* [Granovetter, 1978], dans lesquelles une information se propage au fil du temps dans un *graphe de diffusion*, dont les nœuds sont initialement non-informés ou inactifs ; ils sont activés à réception de l'information, et le demeurent jusqu'à la fin de la simulation de la propagation.

**Definition 3.2.2.** *Graphe de diffusion d'information :* il s'agit d'un graphe G = (U, E), où U est l'ensemble des nœuds-utilisateurs, émetteurs et récepteurs d'information, et E l'ensemble des arcs dirigés, servant de support à la diffusion de l'information.

Dans le modèle des seuils linéaires (LT) [Granovetter, 1978], à chaque nœud $u$ est associé un seuil $\sigma_u \in [0,\ 100]$, qui représente le pourcentage nécessaire de contacts (de nœuds voisins de $u$) pour convaincre $u$ de diffuser une information $i$. Les valeurs classiques de $\sigma_u$ sont de 70% ou 80%. Ce mécanisme de diffusion est calculé itérativement, par pas de temps, jusqu'à trouver un équilibre, l'information $i$ ne se propageant pas davantage. Les arcs peuvent être pondérés par une valeur de facilité de propagation : plus le poids est élevé, plus la relation est influente, convaincante.

Dans le second modèle, dit des cascades (IC) [Goldenberg et al., 2001], ce sont les arcs qui reçoivent des attributs de facilité de propagation. Cette représentation provient d'une simulation calculant l'importance relative de la publicité par rapport au « bouche-à-oreille », lui-même scindé selon la relation entre les individus (liens *forts* ou *faibles* selon l'intensité de la relation sociale). En équation 3.1, les paramètres du modèle sont $\alpha$ la probabilité que le marketing et la publicité persuade un individu, $\beta_w$ la probabilité qu'un individu soit convaincu d'adopter une information par ses contacts faibles (ses connaissances distantes, au nombre de $k$), et $\beta_s$ la probabilité qu'il le soit par son cercle privé (liens forts, au nombre de $m$).

$$p = 1 - (1-\alpha)(1-\beta_w)^k(1-\beta_s)^m \tag{3.1}$$

À chaque pas de temps, des tirages aléatoires ont lieu pour chaque nœud, déterminant son activation (adoption et rediffusion de l'information). La simulation tourne jusqu'à avoir propagé l'information parmi toute la population ; plusieurs point initiaux de diffusion permettent de trouver les meilleurs diffuseurs. Le paramètre $\alpha$, représentant l'impact de la publicité, suggère que l'information ne se propage pas seulement le long des arcs, mais aussi sans frontière via les médias d'information. Ainsi, ce modèle *LT* évalue l'*influence* des modes de diffusion de l'information, en considérant tous les utilisateurs comme également influençables : leur seule différence porterait sur leur position sociale, plus précisément sur le nombre de liens, forts et faibles, qu'ils ont tissé autour d'eux.

Une bonne revue de ces méthodes est proposée par [Guille et al., 2013], selon qui le modèle des cascades, *IC*, est centré sur l'émetteur de l'information, alors que les seuils linéaires *LT* se focalise sur

la réception. En effet, *IC* attribue des forces de propagation aux émetteurs, alors que tous les nœuds récepteurs sont uniformes. De l'autre côté, *LT* différencie les récepteurs par des seuils de persuasion personnalisés.

De nombreuses extensions considèrent d'autres aspects et mécanismes ; sans les parcourir exhaustivement, nous nous penchons sur l'une d'entre elles, qui exploite les profils des utilisateurs et prend en compte les contenus échangés [Lagnier et al., 2013]. Les *profils* correspondent ici aux préférences (centres d'intérêt) des utilisateurs, exprimées comme des vecteurs dans l'espace des *contenus*. De cette façon, une fois atteints par l'information, les utilisateurs ne la propagent que si elle correspond à leurs centres d'intérêt. Chaque nœud est aussi caractérisé par un *rôle*, actif, ou passif : il s'agit de la propension habituelle d'un compte à retransmettre les informations. Enfin, la diffusion elle-même est renommée « pression sociale », dont l'implémentation s'inspire des modèles précédents (*IC* et *LT*).

### 3.2.2 Influenceurs dans un modèle de diffusion

Une première définition d'un « influenceur » est basée sur son importance dans la diffusion et la propagation de l'information : les comptes importants sont ceux qui peuvent orienter le débat voire faire finalement valoir leur point de vue.

Les modèles que nous venons d'introduire peuvent être exploités pour détecter les influenceurs. Souvent utilisé à des fins de marketing ou de publicité, le problème (aussi nommé *Influence Maximization*) se pose comme la sélection d'un ensemble de comptes « influents » afin de diffuser au mieux, c'est-à-dire au plus large sur le réseau social considéré, un message commercial : où commencer la diffusion du nouveau produit ? Comment profiter au mieux des influenceurs déjà en place ? Ici encore, le problème hérite du domaine de l'épidémiologie : une fois « contaminé », le nœud propage l'objet de la simulation, à savoir l'unique information considérée.

Mathématiquement défini par [Kempe et al., 2003], le problème de la sélection de $k$ influenceurs-nœuds du graphe pour propager au mieux une information après un nombre fixé de pas de temps est NP-difficile [3]. La difficulté de trouver la solution exacte est compensée par l'existence de bonnes heuristiques : si la sélection des nœuds de plus haut degré ou centralité est intuitive, la solution gloutonne rend de meilleurs résultats. Il s'agit de l'ajout successif des nœuds contribuant le plus à maximiser la fonction objectif (le nombre de nœuds activés par l'information).

Pour répondre à un cas précis sur des données réelles, il faut représenter le réseau de diffusion par un graphe dont les nœuds disposent d'attributs supplémentaires provenant du média social considéré : par exemple, le nombre de retweets et de mentions. Ainsi, un mécanisme « social » permet de prendre ces valeurs en compte et favoriser la diffusion du message selon la proximité des comptes, c'est-à-dire l'intensité de la diffusion effective des messages précédents, et ainsi mieux sélectionner les influenceurs [Jendoubi et al., 2017]. Cette approche considère les contacts d'un influenceur comme étant eux-mêmes plus influents ainsi liés, qu'isolés : la topologie du graphe des contacts détermine clairement l'influence des individus.

### 3.2.3 Limites des modèles de diffusion

Cette approche à base de modèles de diffusion rencontre plusieurs problèmes : d'une part, l'information peut parfois se propager de manière instantanée, puisque les contenus publiques sont ouverts à la recherche, qui est très populaire sur Twitter. Les tweets sont souvent lus par des personnes non abonnées, mais qui souhaitent se renseigner sur un mot ou un hashtag. De plus, les mêmes utilisateurs sont aussi des consommateurs d'autres médias, d'où peut provenir ou transiter l'information : cela introduit des raccourcis invisibles sur la représentation de la diffusion dans le réseau social. Cet aspect est certes considéré par le modèle des cascades *IC*, via le paramètre $\alpha$ (influence de la publicité) qui peut faire apparaître l'information sans précédence d'un lien dans le graphe de diffusion ; cependant il est souvent ignoré dans les variantes du modèle.

3. Cette affirmation dépend du modèle de diffusion retenu, mais est valide pour *IC* et *LT*.

D'autre part, la diffusion de l'information dépend aussi de sa nature. Par exemple, il apparaît que les rumeurs et faux contenus se propagent différemment, voire plus vite, que des informations vérifiées [Mendoza et al., 2010]. Or il ne suffit pas de le vouloir pour produire un contenu viral : toutes les publicités ne trouvent pas forcément leur public. De même, les opinions et les informations ne se propagent pas de la même façon. L'information doit « seulement » être lue et comprise ; la propagation de l'opinion fait appel à des mécanismes d'argumentation et de persuasion, qui doivent prendre en compte le charisme de l'émetteur, et les sensibilités du récepteur.

Enfin, la nature du réseau sur lequel la diffusion est étudiée n'est pas triviale ; le réseau des abonnements sert souvent de support naturel puisqu'il relie auteurs et lecteurs supposés, mais n'est pas la seule option. Par exemple, le réseau de la *confiance* peut être exploité à cette fin [Mohamadi-Baghmolaei et al., 2015].

## 3.3 L'influence comme position de confiance

Alors que les scores d'influence attribuent une valeur « globale » à un utilisateur (une réputation), une autre approche, que nous nommons « scores de confiance » consiste à attribuer une valeur à la relation entre deux comptes (la confiance). En effet, tous les comptes ne sont pas aussi sensibles au message propagé par un même influenceur, qui dispose donc d'une capacité d'influence spécifique à chacun de ses lecteurs.

**Definition 3.3.1.** *Confiance :* relation dirigée entre deux comptes. L'assurance que le *confiant* a, portant sur la propension d'un autre compte à agir dans le futur en regard d'un certain objectif.

Ce découpage entre *réputation*, en tant que score global, et *confiance* est présent dans plusieurs travaux [Jøsang et al., 2007, Herzig et al., 2009], dans lesquels la confiance va du nœud $i$ vers le nœud $j$, alors que la réputation de $i$ est une unique valeur vue par tous les membres du réseau.

### 3.3.1 Graphes de confiance

L'obtention d'un graphe de confiance n'est pas aisée : en mesurant l'intensité de la relation maintenue par chaque compte envers les autres, exprimée en quantité d'actions accomplies par l'un envers l'autre, un réseau est construit. Il présente une meilleure précision par rapport à des *baselines* de réputation telles que le degré ou le PageRank [Caverlee et al., 2010].

Dans une revue, une liste des propriétés d'un graphe de confiance est proposée, provenant du domaine psychologique [Sherchan et al., 2013]. Ainsi, ce type de graphe est susceptible de montrer les traits suivants :

— spécificité au contexte. Les nœuds se font confiance sur un domaine (l'exemple retenu concerne la confiance accordée à un médecin pour les problèmes de santé, mais pas pour d'autres problèmes) ;

— évolution dynamique. L'intensité de la confiance varie au fil du temps, ainsi qu'au fil des nouveaux ressentis ;

— propagation de la confiance. Il ne s'agit pas exactement de transitivité, cependant les individus ont tendance à accorder un peu de confiance aux personnes qui leur sont recommandées par leurs contacts de confiance ;

— composabilité et non-transitivité. La propagation de la confiance au long des arcs de confiance (composabilité) n'implique pas une confiance aveugle envers les « amis des amis » (non-transitivité). En outre, la plupart des nœuds ne revendiquent qu'un petit nombre de liens de confiance sortants ; il n'y a donc pas d'élargissement systématique des liens de confiance d'un individu ;

— subjectivité. Les biais et préférences personnelles de chacun doivent être prises en compte, et la manière de réagir aux recommandations et événements peut être radicalement différente d'un nœud à l'autre ;

— asymétrie. Les liens de confiance n'ont pas à être réciproques ;

— auto-renforcement. En effet, de bonnes valeurs de confiance ont tendance à s'améliorer au fil du temps, car les individus ont tendance à agir positivement envers ceux en qui leur confiance est placée ;

— sensibilité aux événements. Un unique événement ponctuel peut détruire une relation longue, de manière irréversible.

Cette liste donne les propriétés d'un modèle de confiance, qui permettent de valider la construction d'un graphe de confiance. Un tel graphe semble requis pour détecter les « vrais » acteurs-clés : ceux qui ont la confiance des utilisateurs, et dont le message sera mieux diffusé. Un interlocuteur de confiance est plus convaincant pour propager une opinion ; dans le cas d'une information, le récepteur devient plus facilement un relais.

La récupération des informations de confiance est souvent fastidieuse, et se révèle non-exhaustive. Impossible de demander un score de confiance envers chacun des autres membres du réseau, pour chaque individu. Aussi l'une des tâches du domaine se focalise sur l'inférence des liens, pour les enrichir dans des graphes de confiance [Golbeck and Hendler, 2004].

L'inférence, c'est-à-dire l'application d'un mécanisme de propagation de la confiance dans un voisinage, résulte finalement en l'obtention d'une *proximité* quantifiée entre chaque paire de nœuds du réseau [DuBois et al., 2009]. Cette proximité est ensuite exploitée, en l'occurrence sur deux jeux de données, le *Trust Project* (62 nœuds) et *FilmTrust* (330 nœuds), pour distinguer des groupes d'utilisateurs. Les auteurs comparent les résultats en prenant ou non en compte la symétrie de la confiance : si la relation est asymétrique, à chaque arc $e_{i,j}$ peut répondre un arc $e_{j,i}$, de valeur différente. En conséquence, la quantité d'information initiale nécessaire est nettement plus importante pour donner des résultats intéressants par rapport aux modèles de propagation de l'information.

### 3.3.2 Versions améliorées de la confiance

Deux innovations ont été apportées sur l'information portée par les arcs : plutôt que de porter une simple valeur de confiance, de 0 à 1, il est proposé de considérer le tuple $(t_{ij}, c_{ij})$. La première valeur, $t$ pour *taste*, représente le goût dans le sens d'approbation : est-ce que $i$ approuve les informations émises par $j$, allant de -1 (désaccord total) à 1 (complètement d'accord) ; et $c$ pour *certainty*, la certitude allant de 0 à 1 : est-ce que $i$ considère que $j$ est honnête, intègre ? Dans ce modèle de diffusion de l'information plus fin, la probabilité que $i$ devienne un relais de $j$ est plus forte si $i$ approuve les contenus ($t$ élevé) et croit en l'intégrité de $j$ ($c$ élevé). Prenant en compte à la fois l'approbation, et la méfiance/confiance, il en résulte un modèle de diffusion de l'information plus précis [Gao et al., 2015].

Une seconde distinction est proposée entre deux types de confiance : la *popularité*, plus proche de la réputation que nous avons introduite, et l'*engagement*, c'est-à-dire la confiance que l'utilisateur place en sa communauté, déterminant sa loyauté et l'intensité de ses actions envers le groupe [Nepal et al., 2011].

Certains travaux considèrent la confiance comme thématique : sa valeur évolue selon les thèmes abordés [Wang et al., 2015, Reed and Kadayam, 2017]. L'un de ces travaux propose un nouveau type de réseau social, où les utilisateurs créent des liens d'abonnement en mentionnant, d'une part la confiance qu'ils portent en l'émetteur, et d'autre part la thématique pour laquelle l'ajout est réalisé : X veut lire ce que Y écrit, si c'est sur le thème T [Reed and Kadayam, 2017].

L'autre référence exploite, en sus de la confiance thématique, la notion de *social intimacy* (intimité) entre deux comptes, propice à l'établissement d'une confiance réciproque [Wang et al., 2015]. La confiance entre deux comptes dépend ainsi de leur *statut*, de l'histoire de leur relation (intimité), du contexte (les actions de tiers sur les deux comptes) et des préférences de chacun (thématiques).

### 3.3.3 Limites de l'approche par confiance

La détection de la confiance, surtout à grande échelle, est ardue : d'une part il s'agit de quelque chose d'invisible et subjectif, et d'autre part le dynamisme et le nombre d'acteurs en jeu compliquent le maintien à jour de l'information de confiance entre individus [Jiang et al., 2014]. La plupart des approches présentées ici basent leur validation sur des jeux de données de deux types. Soit il s'agit de donnée générée artificiellement, soit il s'agit d'une compilation de relations de confiance, obtenue en proposant un remplissage de questionnaires individuels : cette méthode n'est appliquée que pour des petites populations (moins de cent individus).

Une vraie détermination des acteurs-clés doit considérer, voire se baser sur un graphe de confiance ; cependant sa construction est excessivement difficile et propice à l'erreur. En contrepartie, la réputation propose une autre piste de réponse au problème posé, et ses effets nous semblent plus aisés à mesurer.

## 3.4 Réputation et position sociale

Plus tangibles que la notion de confiance, les abonnements et échanges entre comptes utilisateurs servent eux aussi de matière première pour construire des graphes. Les nœuds y sont les comptes utilisateurs, reliés selon les interactions qu'ils réalisent. L'influence d'un individu est souvent imaginée comme correspondant à sa position dans le réseau : certaines positions sont plus fortes que d'autres. Cette « force » est mesurée grâce aux indicateurs topologiques suivants. Dans un second temps, il s'agira d'étiqueter des types de positions d'un nœud dans un graphe, et enfin, de prendre en compte les statistiques liées à l'implémentation de la plate-forme analysée.

### 3.4.1 Indicateurs topologiques : centralités, PageRank

Les premières mesures de l'importance d'un nœud dans un graphe se basent sur le nombre de connexions dont dispose un nœud, c'est-à-dire son **degré**. Pour un nœud donné, il faut distinguer le nombre d'arcs dirigés vers (degré entrant), sortant de (degré sortant), ou la somme des deux : ainsi un compte spammeur émettant de nombreuses mentions, qui reste constamment ignoré par ses pairs, affiche un fort degré sortant, mais un faible degré entrant. De nombreux travaux ont mesuré la distribution du degré au sein du graphe, révélant sa répartition inéquitable [Kwak et al., 2010] : à l'opposé d'un graphe aléatoire où une même probabilité uniforme mène à l'ajout d'un arc entre deux nœuds, les réseaux sociaux numériques présentent une distribution proche (mais différente) d'une loi de puissance ou d'une loi de Zipf[4]. Ainsi, par exemple, si le nœud de degré maximal dispose d'un million d'arcs, alors il y a un million de nœuds disposant d'un unique arc.

Le degré ne suffit pas à estimer l'importance d'un nœud dans le graphe, à faire la différence entre le cœur du réseau et une position périphérique. C'est pourquoi plusieurs méthodes se penchent sur la notion de **centralité**. Plusieurs variantes existent (centralités de degré, de proximité, d'intermédiarité, de Katz, ...), mais la plus communément retenue est la *centralité de proximité*, introduite par [Bavelas, 1950]. Elle donne du poids aux nœuds qui sont positionnés le moins loin de tous les autres, en moyenne. Plus précisément, elle est calculée comme l'inverse de la somme des longueurs des plus courts chemins entre un nœud et tous les autres nœuds du graphe ; en conséquence elle favorise les nœuds qui sont à l'intérieur d'un groupe dense.

L'**intermédiarité**, ou centralité d'intermédiarité, évalue l'importance d'un nœud dans le processus de diffusion : plutôt que de distinguer les nœuds centraux, elle distingue les nœuds par lesquels l'information doit forcément passer. Dans des graphes présentant des ensembles denses de nœuds,

4. Du nom d'un linguiste, qui s'intéressait à la fréquence d'emploi des mots `https://fr.wikipedia.org/wiki/Loi_de_Zipf`.

on peut remarquer des « ponts », parfois de degré faible, qui relient deux ensembles (ou plus) : cette mesure sert parfois de base pour détecter des communautés dans un graphe [Newman and Girvan, 2004]. Pour réaliser cette distinction, l'intermédiarité du nœud $x$ est calculée comme le nombre de plus courts chemins entre chaque paire de nœuds du graphe, passant par le nœud $x$ ; en conséquence, le calcul de tous les plus courts chemins est nécessaire, impliquant une complexité plus élevée, en $O(n^3)$ dans le pire des cas.

Introduit par les fondateurs de Google [Page et al., 1999], le **PageRank** favorise les nœuds recommandés par les autres, via un mécanisme de propagation de la recommandation. Ainsi, dans le calcul de l'importance du nœud $x$, l'importance des nœuds émettant des arcs vers le nœud $x$ est prise en compte, compensée par leur degré sortant. Un PageRank simplifié, provenant de [Page et al., 1999], est présenté en équation 3.2 : le score $R(u)$ du nœud $u$ dépend de $c$ un facteur de normalisation, et de $N_v$ le degré sortant du nœud $v$, élément de l'ensemble $B_u$ contenant tous les voisins pointant vers $u$.

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \tag{3.2}$$

Cette équation génère des problèmes pour le calcul, s'il existe un cycle continu de liens sans sortie : par exemple, lorsque $i$ et $j$ sont reliés l'un à l'autre, sans proposer de lien sortant vers un nœud tiers. La solution retenue à cette difficulté consiste en l'ajout d'un terme à l'équation, qui peut être vue comme un score *a priori*, noté $E(u)$ dans l'équation 3.3. Afin d'obtenir une solution unique, la somme des $pr$ pour tous les nœuds du graphe fait 1.

$$pr(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} + c\, E(u) \tag{3.3}$$

Le calcul du PageRank peut se poser comme le calcul du vecteur propre dominant (associé à la plus grande valeur propre) de la matrice d'adjacence du graphe, notée A : l'équation 3.2 peut se réécrire $R = cAR$. Auparavant, une première étape consiste à modifier la matrice A pour la normaliser afin que $A_{i,j}$ contienne la probabilité de choisir l'arc $e_{i,j}$ au départ de $i$ (plutôt que le poids brut de l'arc $e_{i,j}$).

L'algorithme de calcul est itératif, avec une condition d'arrêt lorsque les scores évoluent peu d'une itération à l'autre. Il porte un risque de non-convergence selon son initialisation, aussi une solution initiale correspondant à un score uniforme de $\frac{1}{n}$ est souvent choisie. La valeur numérique du score porte peu de sens, mais l'ordre (*Rank*) est pertinent : les scores les plus élevés sont effectivement les mieux placés, c'est-à-dire qu'ils bénéficient de connexions entrantes depuis des nœuds eux-mêmes bien connectés.

TABLEAU 3.1 – Mesures calculées sur un graphe

| Nom | Notation | Description |
|---|---|---|
| **degré** | $d_x$ | nombre d'arcs reliant le nœud |
| **centralité** | $C(x)$ | proximité de tous les autres nœuds |
| **intermédiarité** | $C_B(x)$ | point de passage obligé |
| **PageRank** | $pr(x)$ | référencé par les références |

La table 3.1 résume les quatre mesures exploitant la topologie des graphes. À chaque nœud $x$ correspond des valeurs plus ou moins grandes et dépendant de l'instance du graphe : en effet, pour deux graphes de même nombre de nœuds et d'arcs mais de topologies différentes, tant la répartition du degré que son diamètre ou sa structure (présence de communautés, c'est-à-dire d'ensemble de nœuds densément reliés), ont des conséquences sur la centralité, l'intermédiarité et le PageRank.

### 3.4.2 Classification non-supervisée des positions dans le réseau : le rôle

Dans un réseau social, les nœuds ne sont pas anonymes ni interchangeables : ils sont caractérisés autant par des informations internes (leur identité), que par leur position dans le graphe : comment ces nœuds sont liés à leur entourage. Pour caractériser cette position, nous venons de voir des méthodes de calcul de l'importance d'un nœud dans un graphe, ou plutôt de l'importance de la position d'un nœud dans le graphe. Dans cette section, nous abordons les méthodes d'étiquetage du **rôle** : le « type » de position occupée par un nœud.

En effet, les caractéristiques topologiques, qui décrivent la position d'un nœud dans le réseau, peuvent alimenter un modèle d'apprentissage non-supervisé chargé d'attribuer des rôles aux nœuds. L'intuition consiste à grouper les comptes par caractéristiques similaires, enrichissant l'analyse de la population par cette dimension topologique dans un graphe social. Cette position-type est appelée *rôle* dans les travaux que nous analysons ci-dessous, qui exploitent tous deux une méthode de sélection de caractéristiques [Henderson et al., 2011], parmi lesquelles :

— mesures « locales », qui incluent les degrés entrant, sortant et total, en nombre, et en poids d'arcs, ainsi que la centralité de proximité des voisins ;

— mesures issues des *ego-networks* [5], qui reposent sur les degrés maximum et moyen des voisins, et l'intermédiarité du nœud dans son ego-network.

Un modèle théorique propose de réaliser la classification non-supervisée grâce à une décomposition matricielle [Henderson et al., 2012]. Une matrice V (pour *vertices*) de taille $n \times f$ contient les $f$ caractéristiques de chacun des $n$ nœud. L'approche retenue pour apprendre à la fois les rôles, et les attributions de chaque nœud aux rôles, consiste à décomposer la matrice V pour obtenir, tout d'abord, une matrice G de taille $n \times r$, représentant la similarité d'un nœud à un rôle. Ainsi, $G_{i,j}$ représente à quel point le nœud $i$ correspond au rôle $r_j$. D'autre part, la décomposition de V fournit la définition des rôles dans une matrice F de taille $r \times f$, décrivant les rôles via les caractéristiques retenues. Une infinité de matrices G et F permet d'approximer V, aussi une condition supplémentaire définit le problème mathématique : il faut trouver les matrices G et F dont tous les termes sont positifs ou nuls, et de norme minimale. Le choix de la norme $L_1$ par exemple, permet d'assimiler les termes de G comme des pourcentages d'appartenance d'un rôle aux nœuds.

L'algorithme *RolX* [Henderson et al., 2012], qui calcule les caractéristiques topologiques puis utilise une factorisation matricielle non-négative (NMF), revendique une complexité en $O(mf + nfr)$, où $n$ nœuds liés par $m$ arcs sont décrits par $f$ caractéristiques et résumés par $r$ rôles.

Des travaux supplémentaires proposent d'affiner la décomposition matricielle afin que G et F soient aussi creuses que possibles (*sparsity*), et d'ouvrir la possibilité d'ajouter des conditions, c'est-à-dire de guider la découverte des rôles (*guidance*). Ainsi, un système nommé GLRD (*Guided Learning for role discovery*) [Gilpin et al., 2013], inclut la possibilité de restreindre la quantité de caractéristiques descriptives initiales, d'affilier chaque nœud à un nombre limité de rôles, ou de forcer l'attribution d'un rôle similaire pour un ensemble de nœuds. Appliqué à un réseau de citations dans des publications scientifiques, GLRD identifie quatre rôles-types : les périphériques isolés, les périphériques groupés (petite communauté relativement isolée du reste du réseau), le cœur du réseau, et enfin les « stars » dont le degré est élevé.

Les catégories obtenues dépendent de l'instance de graphe analysé, et peuvent différer suite à l'ajout de quelques arcs. De plus, la découverte de rôles exploite des modèles où l'annotation par un expert est ardue. Comme le rôle correspond au type de position dans un graphe donné, l'expert doit connaître à la fois l'individu, et le graphe. Sans vérité terrain, ce domaine propose une validation par l'explication des résultats non-supervisés qu'il propose ; cette explication est facilitée lorsqu'un lien est disponible entre l'activité (la production de tweets) et le rôle des nœuds-utilisateurs.

5. Sous-graphe centré sur un nœud, contenant par exemple tous les nœuds voisins (profondeur 1), et leurs voisins (profondeur 2), etc.

### 3.4.3 Statistiques liées à l'implémentation du réseau social numérique

L'influence est souvent vue comme la capacité d'un utilisateur à « engager » les autres, c'est-à-dire à susciter leur réaction : typiquement, déclencher un retweet ou retenir leur attention et les faire lire le lien proposé. En l'occurrence, les réseaux sociaux numériques proposent déjà des indicateurs correspondant à l'utilisation faite des fonctionnalités implémentées : sur Twitter, ces statistiques incluent le nombre d'abonnés (*followers*) et de retweets, qui accompagnent respectivement les comptes et les messages.

Pour déterminer qui sont les acteurs influents du réseau, une approche consiste à pré-définir des catégories de comptes, et d'attribuer des étiquettes d'influenceurs ou non. Un tel modèle comprenant cinq classes a été introduit par [Chen et al., 2014a] ; l'annotation humaine est réalisée à partir de la page du compte (incluant les informations du compte et les tweets les plus récents). Les cinq classes sont les suivantes :

— fans, c'est-à-dire les supporteurs inconditionnels d'une marque, de produits ;

— disséminateurs d'informations, dont principalement les médias et blogs d'information ;

— experts, qui sont restreints à un domaine, et se focalisent sur les informations techniques ;

— célébrités ou personnalités publiques dotées d'une grande audience, elles ont un grand impact sur la diffusion d'un message ;

— autres, c'est-à-dire les comptes sans influence particulière.

La structure de données représentant les comptes utilisateurs comporte trois aspects : la thématique (par un vecteur $tf.idf$), le sentiment (par les polarités mesurées dans les tweets) et la popularité (par le nombre d'abonnés, d'abonnements, et un booléen si le compte est vérifié par Twitter). Les auteurs entraînent ensuite un classifieur pour catégoriser automatiquement les influenceurs. Cependant, ce découpage correspond à une application purement publicitaire : outre la faiblesse des scores de classification obtenus, cette taxonomie considère plus le type (ici : fan, média, célébrité...) que le rôle (par exemple émetteur, relais, récepteur) de l'utilisateur.

Pour aider à comprendre ce qu'est Twitter, une comparaison d'indicateurs numériques (le nombre de retweets, d'abonnés, et le PageRank issu du graphe des abonnements) a été proposée dans plusieurs travaux [Kwak et al., 2010, Lee et al., 2010] : les distributions de nombre d'abonnés, de tweets émis et de retweets reçus permettent de quantifier l'usage que les utilisateurs font de la plate-forme. Ces travaux comparent les statistiques proposées par Twitter, ou découlant directement des fonctionnalités implémentées (par exemple, le nombre de retweets), et des scores plus complexes, calculés sur le graphe des abonnements, tels que la centralité. Le $Top20$ des comptes selon chacun des indicateurs est proposé et commenté, permettant à chacun de se faire une idée de l'aspect mis en avant, telles que la célébrité, la persuasion ou la polémique.

Les mesures topologiques (degrés, centralités, PageRank) nécessitent un support, un graphe, pour être calculées. L'exploitation des liens d'abonnement aboutit à la construction d'un graphe reliant les utilisateurs, les plaçant les uns par rapport aux autres. Dans la littérature, un système utilise uniquement le PageRank, et établit la liste des comptes les plus influents de manière quotidienne [Noordhuis et al., 2010]. Les liens d'abonnement sont continuellement demandés auprès de Twitter afin de conserver le graphe à jour ; il y a malgré tout un décalage entre le graphe et la réalité, du fait de la vitesse de récupération de l'information. De plus, le score dépend ici uniquement des abonnements, sans relation avec les actions et réactions suscitées.

La lecture brute des nombres d'abonnés et de retweets est souvent critiquée, car elle n'admet aucune subtilité : recevoir une centaine de mentions n'est pas un événement pour un compte qui dispose de milliers d'abonnés. Pour contrer cela, un score d'influence est calculé, dépendant de l'*affiliation* (proportion d'abonnés réciproques par rapport à l'ensemble des abonnements et abonnés) et de l'*interest rate* (défini pour l'occasion comme le nombre total de mentions d'un compte) [Zam-

paras et al., 2015].

L'un des « scores d'influence » les plus reconnus sur le marché de l'influence marketing online se nomme *Klout*. Il a été conçu pour agréger la présence d'un utilisateur sur plusieurs plate-formes (Facebook, Twitter, Wikipedia, Youtube, autres) [Rao et al., 2015]. Tous les indicateurs récupérés sur ces plate-formes sont rassemblés en *Activités*, *Profils* et *Graphe*. Les *Profils* permettent de mieux caractériser les utilisateurs : experts sur un domaine, position géographique revendiquée, recommandations. Le *Graphe* ne concerne que Wikipedia, et permet de calculer l'importance d'une personnalité ou institution par son score PageRank sur le graphe des pages de l'encyclopédie collaborative. Le cœur du score *Klout* se base sur les *Activités* représentant les réactions suscitées par un utilisateur présent sur plusieurs plate-formes sur une longue période (la durée retenue est de 90 jours).

Le calcul de ce score, dont la formule n'est pas explicitement fournie par ses auteurs, prend en compte la quantité de réactions suscitées (vues sur Youtube, commentaires sur toutes les plate-formes, retweets sur Twitter, ...), de même que la quantité de comptes ayant réagi et leurs propres scores. Ainsi, une différence est faite entre 100 retweets émis par un ensemble de 10, ou de 50 comptes de même influence individuelle. Dans le premier cas, l'influenceur a totalement convaincu son audience, de taille limitée. Dans le second cas, il a moyennement convaincu une population cinq fois plus large.

Cette distinction prend forme par l'usage d'une pondération, déterminée par apprentissage automatique. Cette étape a nécessité la constitution d'un corpus, où les annotateurs devaient choisir lequel de deux comptes était le plus influent ; environ un million de comparaisons entre paires de profils y ont été effectuées. Des éléments de validation du modèle sont proposés : comparaison des classements des scores *Klout* avec le classement des meilleurs joueurs de tennis, ou des femmes les plus influentes selon *Forbes*, ou encore avec le score *Google Trend* ; tous montrent une corrélation forte vis-à-vis de *Klout*.

### 3.4.4 Limites de la réputation

Pour classer les utilisateurs par ordre d'influence, la littérature propose plusieurs approches : des scores topologiques, souvent calculés sur le graphe des abonnements, constituent la majorité des propositions. Cet indicateur de la qualité de la position occupée est complété, parfois validé par les nombres d'abonnés et de messages cités. Ces quantités brutes sont parfois difficiles à comparer entre deux périodes ou thématiques différentes, car les nombres typiques d'abonnés diffèrent selon la popularité du réseau pour la période, langue et mots-clés retenus pour la collecte des données.

Dans ces travaux, le focus est mis sur l'élaboration d'un unique score, pondérant toutes les façons d'être influent. L'influence elle-même y est rarement définie et les validations sont réalisées en comparaison avec des classements d'acteurs-clés dans la vie réelle. Pour notre part, nous pensons qu'un indicateur unique est certes séduisant, mais ne peut pas couvrir avec subtilité l'étendue du problème.

La catégorisation automatique des comptes fait ressortir des acteurs influents mais dépend à la fois de la taxonomie fournie, adaptée à une lecture du problème, et de la représentation des comptes retenue. L'apprentissage non-supervisé, à partir des caractéristiques topologiques de chacun des nœuds d'un graphe, permet d'éviter ces deux écueils : bien qu'elle ne fournisse pas de score d'influence, la détection de rôle nous semble une approche intéressante pour explorer la population de comptes. Les différents rôles ouvrent une nouvelle dimension d'analyse, complétant la vue alimentée par l'étude du comportement et les scores d'influence précédemment introduits.

## 3.5 Synthèse et discussion sur l'analyse des comptes utilisateurs

Dans ce chapitre, nous avons proposé une revue de la littérature dans le domaine de la caractérisation des comptes d'utilisateurs de réseaux sociaux numériques. En effet, la lecture des médias sociaux nécessite de connaître l'émetteur autant que de comprendre le message. En table 3.2, nous proposons une grille de lecture du domaine, distinguant les *modèles* théoriques, les propositions de *mesures* ou

de scores, et les méthodes de *catégorisation* (*i.e.*, d'étiquetage).

La connaissance d'un compte passe par la constitution d'un **profil**, une structure de données relativement complète qui rassemble les informations sur l'émetteur de messages. Les travaux existants compilent différents indicateurs, portant tant sur l'aspect déclaratif (nom, profession et lieu revendiqués) que sur le comportement (l'enchaînement temporel des actions accomplies par un compte).

Pour analyser les utilisateurs, les approches sont multiples : la construction de profils psychologiques, avec le modèle OCEAN [Costa Jr et al., 1991] appliqué sur des comptes Twitter [Quercia et al., 2011] ; le besoin de manipuler l'entité « compte », pour identifier l'individu et relier une présence sur plusieurs plate-formes [Liu et al., 2014], ou pour identifier les comptes malicieux [Subrahmanian et al., 2016] ; l'analyse temporelle d'un profil, pour prédire les périodes de connexion [Gatti et al., 2013] ou pour identifier des comportements-types [Raghavan et al., 2014].

La mesure de l'**influence** se décline au long de plusieurs approches : les modèles de diffusion [Granovetter, 1978, Goldenberg et al., 2001] posent les fondations nécessaires pour identifier les meilleurs (re-)diffuseurs de l'information [Noordhuis et al., 2010, Guille and Favre, 2015]. La même philosophie, appliquée sur la notion de confiance [Sherchan et al., 2013], aboutit à des résultats similaires [Caverlee et al., 2010]. Enfin, la classification en influenceurs ou non [Chen et al., 2014a] mais surtout l'élaboration d'un score d'influence [Kwak et al., 2010, Lee et al., 2010, Rao et al., 2015], souvent à base d'analyse de l'importance d'un nœud dans un graphe, nous amènent à considérer autrement la position sociale : elle constitue une bonne donnée d'entrée pour catégoriser le rôle adopté par le compte [Henderson et al., 2012, Gilpin et al., 2013].

L'influence est un terme qui recouvre plusieurs notions parmi lesquelles nous retenons deux choses distinctes couvrant, selon nous, la majeure partie du concept : la mesure de la position sociale de l'individu, ainsi que la mesure de sa capacité à susciter la réaction. Ces deux éléments sont effectivement mesurables, concernent deux aspects très différents de l'influence, et sont informatifs. La fonction de *mesure* est complétée par la catégorisation des positions dans le graphe, avec les *rôles*.

TABLEAU 3.2 – Récapitulatif des travaux présentés

| | Approche | Modèles | Mesures ou Scores | Catégorisation |
|---|---|---|---|---|
| Profil | Psychologique | [Costa Jr et al., 1991] | [Quercia et al., 2011] | |
| | Identité | | [Liu et al., 2014] | [Subrahmanian et al., 2016] |
| | Activité | | [Gatti et al., 2013] | [Raghavan et al., 2014] |
| Influence | Diffusion | [Granovetter, 1978]<br>[Goldenberg et al., 2001] | [Guille and Favre, 2015]<br>[Noordhuis et al., 2010] | |
| | Confiance | [Sherchan et al., 2013] | [Caverlee et al., 2010] | |
| | Réputation | | [Kwak et al., 2010]<br>[Lee et al., 2010]<br>[Rao et al., 2015] | [Chen et al., 2014a]<br>[Henderson et al., 2012]<br>[Gilpin et al., 2013] |

L'étude des acteurs présents sur un réseau social présente plusieurs défis : il faut mesurer et distinguer les comptes les uns des autres, selon leur comportement, leur impact et leur position. Ces notions sont nécessaires pour prendre en compte les différentes manières d'être et d'agir en ligne. Nous pensons qu'il faut les combiner pour explorer efficacement les médias sociaux.

Dans l'état de l'art, ces notions sont toujours distinctes, et parfois trop spécialisées. L'analyse du comportement englobe plusieurs aspects, chacun traité isolément pour une application donnée ; la mesure de l'influence prend en compte tour à tour la taille de l'audience, la capacité à diffuser, ou la quantité de réactions suscitées ; l'étude de la position dans un graphe reliant les comptes se base sur des liens flous (la confiance) ou sur une relation faible (l'abonnement n'implique pas la lecture des publications).

Dans le chapitre 6, nous proposerons un modèle comprenant les trois notions vues : la construction

d'un profil décrivant le comportement, aspect par aspect, l'élaboration d'indicateurs d'influence fins, et l'étude de la position sociale des comptes utilisateurs.

Basé sur l'entité « compte utilisateur », ce domaine ne tire pas complètement profit de la richesse des graphes sociaux, où l'on voit émerger des structures sociales, des groupes d'utilisateurs qui interagissent entre eux, et agissent sur le reste du réseau social ; nous abordons cette problématique dans le chapitre 4.

CHAPITRE 4

# Détection de communautés par l'analyse de graphes relationnels

L'exploration des médias sociaux révèle l'émergence d'une structure, facilitant ou ralentissant la propagation de l'information selon les liens tissés entre les comptes utilisateurs. Au gré des échanges, des ensembles de comptes se relient fortement, agissant ensemble autour d'un même sujet, d'une même thématique : ces ensembles sont souvent appelés des *communautés* dans la littérature.

Ces communautés se matérialisent de plusieurs façons : par la similarité des profils, des personnalités suivies ou des échanges, ou bien par les liens directs et forts entre les comptes. Ce second type de communauté retient notre attention ; l'analyse des graphes de relations et des interactions entre comptes est l'approche la plus fréquemment explorée afin d'identifier ces groupes d'utilisateurs. La pertinence de ces groupes est alors mesurée grâce à des indicateurs topologiques, évaluant par exemple la quantité de liens internes au groupe.

Dans ce chapitre, nous proposons une revue de la tâche de détection et de caractérisation des communautés à partir de graphes et de réseaux sociaux. Dans un premier temps, la notion de communauté est exposée en section 4.1, puis les définitions et mesures qui évaluent la force des liens entre les membres d'une même communauté sont présentées en section 4.2. Ensuite, les algorithmes de détection de communautés sont introduits et comparés en section 4.3. Quelques applications de ces algorithmes et mesures à des données réelles, à la recherche de communautés thématiques, sont présentées en section 4.4. Finalement une synthèse, en section 4.5, conclut ce chapitre.

## 4.1 Notion de communauté

La notion de communauté correspond à une même intuition dans deux domaines différents : de manière large, il s'agit de groupes d'individus reliés par leurs similarités ou leurs relations, des individus qui ont quelque chose *en commun*. Nous précisons cette notion dans un premier temps en analyse des médias sociaux, où il existe plusieurs supports permettant de délimiter les communautés ; puis nous abordons l'aspect informatique et mathématique du problème, en analyse de graphe ou de réseaux.

### 4.1.1 En analyse de médias sociaux

Certaines plate-formes (Facebook par exemple) incluent une fonctionnalité de création de groupes, autorisant des utilisateurs à échanger sur une page en commun. Cependant, tous les réseaux ne proposent pas cette fonctionnalité : c'est encore le cas de Twitter actuellement. Que la fonctionnalité existe ou non, la lecture par groupes d'utilisateurs ou par communautés éclaire la dimension sociale

de l'analyse des médias sociaux. Selon l'objectif de l'analyse menée, les communautés correspondent à des dimensions très différentes dans la littérature. Il peut s'agir de communautés :

- thématiques, correspondant par exemple à un ensemble de comptes qui mentionnent un même mot-clé, une même notion, ou qui sont abonnés à un même compte ;
- sociales, tel un ensemble de comptes liés entre eux (par abonnements ou interactions réciproques, par exemple les amis d'un compte d'intérêt) ;
- spatiales, c'est-à-dire un ensemble de comptes provenant d'une même aire géographique ;
- catégorielles, *id est* un ensemble de comptes partageant une similarité (âge, catégorie socio-professionnelle supposée ou revendiquée) ;
- linguistiques, comme un ensemble de comptes utilisant la même langue, le même dialecte ou jargon.

Nous donnons ci-dessous plusieurs exemples de travaux basés sur ces définitions de communauté : un premier exemple cherche à visualiser l'activité politique pré-électorale sur Twitter. Pour ce faire, une extraction de sentiment appliquée sur les tweets donne des courbes d'évolution dans le temps. Il faut encore rattacher ce sentiment à des notions, des cibles d'opinions, ou des hashtags : les co-occurrences des hashtags alimentent un algorithme de détection de communauté, permettant aux auteurs de rattacher les clusters de mots-clés à des sensibilités politiques, puis d'analyser la popularité et le sentiment de chacune de ces thématiques et sensibilités. Cette méthode est appliquée sur plusieurs jeux de données, chacun représentant un mois de tweets (campagne électorale italienne pour les législatives de 2014). Le nombre d'utilisateurs produisant des messages dans le même cluster thématique, mois après mois, est conservé pour évaluer l'évolution du débat politique [Amelio and Pizzuti, 2015].

Dans une autre expérience, une communauté est définie comme un groupe de comptes s'exprimant à propos du même thème [Doan et al., 2006] : tout participant à une conférence académique sur l'informatique, fait partie de la communauté des chercheurs en informatique.

Exploitant les relations d'abonnement, dans le jeu de données mis en ligne par Kwak [Kwak et al., 2010], et depuis retiré à la demande de Twitter, un modèle d'identification de communautés « d'intérêts communs » est construit sur la base de célébrités représentatives de ces « intérêts communs », qui incluent des notions larges (par exemple *Film & TV*, *Music*, *Politics*) [Lim and Datta, 2012]. Ainsi deux comptes partagent un même intérêt s'ils sont tous deux abonnés à la même célébrité. Dans une seconde étape, le graphe des abonnements reliant des utilisateurs tous abonnés aux mêmes célébrités est ingéré par un algorithme de détection de communautés (CPM, *clique percolation method*, [Palla et al., 2005]), résultant en des groupes partageant forcément les mêmes centres d'intérêt.

À partir des textes de tweets, une méthode de regroupement sémantique permet d'obtenir des groupes de comptes s'exprimant de manière similaire, et de calculer un score de *tension* entre communautés [Burnap et al., 2015]. Cette tension est une dissimilarité sémantique : calculée sur le texte, elle prend notamment en compte la présence d'injures, d'argot (*slang*). Cette méthode constitue le socle de COSMOS, une plate-forme de supervision des réseaux sociaux, fournie à la presse et aux médias. Selon ces travaux, les groupes ne présentent pas forcément de liaisons entre leurs membres : aucune condition d'existence d'interaction entre les comptes n'est requise, le groupe n'existe « que » par l'émission de messages similaires, de même thématique.

Ces approches diverses exploitent peu la notion de graphe. Il s'agit pourtant d'un domaine aux solides fondations théoriques, au sein duquel le concept de communauté est défini et mesurable.

### 4.1.2 En analyse de réseaux

Une intuition globalement répandue consiste à modéliser un réseau social par un graphe, $G = (V, E)$, dont les nœuds $i \in V$ sont les comptes utilisateurs, reliés par des arcs $e \in E$ représentants

leurs relations : abonnement, amitié, envoi de message, mention, réponse. Il est possible de voir ces graphes par les communautés qui les composent, dont la définition donnée par Newman est la suivante : [*communities are*] « *densely connected groups of vertices, with only sparser connections between groups* » [Newman, 2006]. Nous la traduisons en définition 4.1.1.

**Definition 4.1.1.** *Communauté :* groupe ou ensemble densément connecté de nœuds dans un graphe, présentant des connections moindres vers d'autres groupes.

Tous les graphes ne présentent pas forcément des communautés aussi denses en interne, et isolées les unes des autres. Afin de clarifier ce point, nous citons deux modèles de génération aléatoire de graphes. Le premier modèle, introduit en 1959 par Erdös et Rényi [Erdös and Rényi, 1959], propose de fixer $n$ le nombre de nœuds, et de choisir $p$ la probabilité d'existence de chacun des arcs. Pour chaque paire de nœuds $i, j$, un tirage aléatoire détermine si l'arc $e_{i,j}$ existe. En conséquence, le degré d'un nœud suit une loi binomiale, et la distribution du degré de tous les nœuds du graphe suit une loi de Poisson [Newman et al., 2001].

Cette situation diffère grandement de ce qui est constaté habituellement sur les réseaux sociaux : sur le graphe des abonnements de Twitter, ou sur le graphe des amitiés de Facebook, la distribution est plus proche d'une loi de puissance. La proportion $p(k)$ de nœuds de degré $k$ est proportionnelle à $k^{-\gamma}$, où $\gamma$ est un paramètre propre au réseau (souvent, entre 2 et 3) [Onnela et al., 2007], auquel cas, le graphe est qualifié de réseau invariant d'échelle (*scale-free network*). Ce type de réseau, dont la distribution est moins équilibrée que dans les graphes aléatoires, favorise l'apparition de « *hubs* », nœuds de fort degré, cœurs de communautés, et de « *bridges* », des ponts entre groupes plus distants.

Une autre différence cruciale repose dans l'indépendance des tirages aléatoires des arcs dans le modèle d'Erdös et Rényi : il n'y a pas de mécanisme correspondant au phénomène « les amis de mes amis sont mes amis ». Cela se mesure par le coefficient de clustering CC(S) (ou coefficient d'agglomération) d'un ensemble S de nœuds, défini en équation 4.1. Il est communément calculé pour un graphe (S = V) ou pour un sous-graphe (S ⊂ V). Un triplet est un ensemble connexe de trois nœuds (de deux, ou trois arcs) ; un triangle (« triplet fermé ») est une clique de trois nœuds (connexe, trois arcs). Une version différente de l'équation compte le nombre de nœuds appartenant à un triangle. Les valeurs prises par C selon le graphe montrent une tendance nette : des valeurs élevées pour des réseaux « réels », plus faibles pour les graphes générés aléatoirement [Watts and Strogatz, 1998].

$$\mathrm{CC(S)} = \frac{\#\{triangles\}}{\#\{triplets\}} \tag{4.1}$$

Une dernière différence notable entre graphes aléatoires et réseaux sociaux réels est basée sur le diamètre du réseau : quelle est la longueur du plus court chemin entre n'importe quelle paire de nœuds ? L'existence de nœuds concentrant de très nombreuses connexions diminue fortement cette valeur : en 2011, un utilisateur de Facebook peut joindre plus de 95% des comptes en un maximum de 6 sauts. Ce phénomène est appelé « réseau petit monde » (*smallworld network*) [Watts and Strogatz, 1998] : en moyenne, la distance entre deux nœuds choisis aléatoirement est proportionnelle à $\log n$, le logarithme du nombre de nœuds du réseau.

### 4.1.3 Partitions et couvertures

Cette section aborde l'état de l'art des méthodes de *détection* qui, ingérant un *graphe* et d'éventuels paramètres, proposent un ensemble de communautés $\Gamma$ (*community set* en anglais) : il s'agit de l'attribution de chaque nœud à une ou plusieurs communautés. Dans ce dernier cas, un coefficient d'appartenance d'un nœud à un groupe est nécessaire.

L'approche la plus classique pour la détection de communautés consiste à attribuer un seul label de groupe pour chaque nœud. De nombreux algorithmes de détection de communautés produisent ainsi une partition d'un graphe, définie en 4.1.2.

**Definition 4.1.2.** *Partition :* découpage d'un graphe G = (V, E) en un ensemble de communautés sans intersection entre elles. L'ensemble $\Gamma$ des communautés $c_1$, $c_2$... $c_k$ est une partition si :

— $\cup_{i=1..k}\ c_i\ =\ \mathrm{V}$ : chaque nœud appartient à une communauté ;

— $\forall i, j\ ;\ i \neq j\ ;\ c_i \cap c_j\ =\ \emptyset$ : l'intersection de deux communautés est vide.

Une seconde approche autorise un nœud à prendre part à plusieurs communautés. Pour les algorithmes de cette famille, l'ensemble $\Gamma$ est alors nommé *couverture*, en définition 4.1.3. Cette notion englobe les partitions. La plupart des algorithmes attribuent chaque nœud à au moins un groupe, et proposent un poids d'appartenance $b_{i,c_j}$ du nœud $i$ au groupe $c_j$. Pour un même nœud, la somme des poids d'appartenance est égale à 1. Cette convention est souvent respectée, mais pas nécessairement.

**Definition 4.1.3.** *Couverture :* découpage d'un graphe $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ en un ensemble $\Gamma$ des communautés $c_1,\ c_2 ...\ c_k$.

Dans la suite de ce chapitre, nous nous concentrons sur les mesures évaluant le nombre de liaisons internes à ces communautés, puis sur les méthodes de détection de communautés à partir de graphes : il ne suffit pas d'avoir des nœuds connexes pour les considérer comme un groupe fortement lié.

## 4.2 Mesures d'évaluation et de comparaison de communautés

Cette section présente deux familles de mesures d'évaluation de la qualité des communautés : en premier lieu, les méthodes topologiques qui calculent et caractérisent les communautés ; en second lieu, des mesures de comparaison de deux ensembles de communautés détectées, qui mesurent la similarité entre deux partitions par exemple.

### 4.2.1 Mesures topologiques

Les caractéristiques topologiques des graphes, telles que la valeur de modularité ou la densité interne des groupes détectés, permettent d'évaluer la qualité des algorithmes de détection de communautés, qu'il s'agisse de partitions ou de couvertures. Une revue très complète des fonctions de scores est proposée dans [Yang and Leskovec, 2015], dont les définitions et formalismes sont repris.

Un graphe $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ contient $n = |\mathrm{V}|$ nœuds et $m = |\mathrm{E}|$ arcs. Une communauté est représentée par S, un ensemble de $n_\mathrm{S} = |\mathrm{S}|$ nœuds, liés par $m_\mathrm{S}$ arcs internes (dont les deux extrémités sont dans S) et $c_\mathrm{S}$ arcs externes (dont seulement l'une des extrémités est dans S).

$$m_\mathrm{S} = |\{(u, v) \in \mathrm{E} : u \in \mathrm{S}, v \in \mathrm{S}\}| \tag{4.2}$$

$$c_\mathrm{S} = |\{(u, v) \in \mathrm{S} : u \in \mathrm{S}, v \notin \mathrm{S}\}| \tag{4.3}$$

La **densité interne (d)** quantifie la proportion d'arcs au sein d'un ensemble de nœuds : c'est le ratio entre le nombre réel d'arcs, sur le nombre maximal possible. Elle prend des valeurs réelles entre 0 : aucun arc interne, et 1 : groupe très densément lié. L'idée qu'une communauté est plus dense qu'un ensemble aléatoire de nœuds, est très répandue. Cependant, pour de grandes communautés, les membres ne sont pas tous reliés les uns aux autres (c'est-à-dire, il ne s'agit pas d'une clique), et cette valeur de densité est alors très faible. Deux formules existent : en équation 4.4 pour les graphes non dirigés, et en équation 4.5 lorsque le graphe est dirigé (deux fois plus d'arcs possibles).

$$d(\mathrm{S}) = \frac{2m_\mathrm{S}}{n_\mathrm{S}(n_\mathrm{S}-1)} \tag{4.4}$$

$$d(\mathrm{S}) = \frac{m_\mathrm{S}}{n_\mathrm{S}(n_\mathrm{S}-1)} \tag{4.5}$$

Le ratio d'appartenance à une triade, ou **Triad participation ratio** (TPR), évalue le nombre de nœuds appartenant à une triade, ou triangle, dans la communauté S. Il est défini en équation 4.6. Une valeur de 1 signifie que le groupe est fortement lié en interne : chaque membre y est lié à au moins deux autres membres, eux-mêmes directement liés entre eux. Ce ratio différencie les communautés

où de nombreux nœuds isolés sont reliés uniquement à un nœud central, en leur attribuant un score faible. Là où le coefficient de clustering, en équation 4.1, quantifiait la proportion de triangles, le TPR se penche sur la proportion de nœuds appartenant à des triangles ; cependant il s'agit de deux indicateurs différents.

$$\mathrm{TPR(S)} = \frac{|\{u, \{(v,w) : (u,v) \in \mathrm{E}, (u,w) \in \mathrm{E}, (v,w) \in \mathrm{E}\} \neq \emptyset\}|}{n_\mathrm{S}} \tag{4.6}$$

La **Conductance** (C) quantifie la proportion d'arcs qui sont dirigés vers d'autres communautés. Elle a été introduite par Shi [Shi and Malik, 2000]. Les valeurs vont de 0 pour un groupe isolé, à 1 pour un groupe très connecté vers le reste du réseau. Une valeur élevée signifie que le groupe ne présente que très peu de liaisons internes ; une valeur faible suggère que le groupe n'a pas de relation, d'influence ou d'impact sur le reste du réseau.

$$\mathrm{C(S)} = \frac{c_\mathrm{S}}{2m_\mathrm{S} + c_\mathrm{S}} \tag{4.7}$$

La **Modularité (Q)** se focalise sur les nombres d'arcs internes aux groupes, par comparaison avec un modèle de graphe plaçant les arcs aléatoirement. Introduite par Newman [Newman and Girvan, 2004], une forte modularité Q, proche de 1, signale une bonne partition au niveau du graphe : il s'agit d'une mesure globale, de qualité d'une partition, par opposition aux mesures locales, de qualité d'une communauté.

L'équation 4.8 expose la formule [1] pour calculer la modularité Q, où A est la matrice d'adjacence du graphe : $\mathrm{A}_{ij}$ est le poids de l'arc allant de $i$ à $j$, $k_i$ est la somme des poids des arcs adjacents à $i$ ; enfin $\delta$ vaut 1 si ses arguments sont égaux, et 0 sinon. Par simplicité, l'équation n'inclut pas l'ensemble des arguments : $\mathrm{Q(G,\Gamma)}$ est la modularité de la partition $\Gamma$ (contenant les $c_i$, information d'appartenance du nœud $i$ à une communauté) appliquée au graphe G (contenant la matrice d'adjacence A, et les degrés des nœuds $k_i$). Deux valeurs de modularité ne sont comparables que lorsqu'elles sont calculées sur le même graphe ; l'ajout ou le retrait de quelques arcs peut résulter en des partitionnements et des modularités complètement différentes.

$$\mathrm{Q} = \frac{1}{2m} \sum_{ij} \left[ \mathrm{A}_{ij} - \frac{k_i * k_j}{2m} \right] \delta(c_i, c_j) \tag{4.8}$$

La modularité est étendue aux couvertures sous la notation $\mathrm{Q}_{ov}$, où $ov$ signifie *overlap* [Nicosia et al., 2008]. En équation 4.9, $k_i^{out}$ (resp, $in$) est le degré sortant (resp., entrant) du nœud $i$. À chaque nœud sont attribués des coefficients $\alpha_\mathrm{S}$ d'appartenance à la communauté S, elle-même élément de la couverture C ; le choix de la méthode de combinaison des poids $\alpha$ pour créer $\beta_{l(i,j),\mathrm{S}}$, le coefficient d'appartenance de l'arc $i, j$ à la communauté S, reste ouvert ; cependant il est recommandé d'opter soit pour la moyenne des $\alpha$, soit pour le maximum.

$$\mathrm{Q}_{ov} = \frac{1}{m} \sum_{\mathrm{S} \in \mathrm{C}} \sum_{i,j \in \mathrm{V}} \left[ \beta_{l(i,j),\mathrm{S}} \mathrm{A}_{i,j} - \frac{\beta^{out}_{l(i,j),\mathrm{S}} k_i^{out} \beta^{in}_{l(i,j),\mathrm{S}} k_i^{in}}{m} \right] \tag{4.9}$$

Pour chaque communauté, ces mesures donnent des clés de lecture de leur qualité et de leurs caractéristiques, telles que l'intensité des liens internes, la relation avec l'extérieur ou encore approchant la proportion de nœuds-feuilles, c'est-à-dire reliés uniquement à un seul autre membre (et donc, hors triangles). N'illustrant pas les mêmes aspects, c'est leur exploitation conjointe qui fournit la meilleure information sur le type de communautés obtenues : deux communautés de même densité peuvent présenter des TPR très différents ; deux communautés directement reliées l'une à l'autre n'attribuent pas la même importance à cette liaison, importance qui est évaluée par leurs conductances respectives.

1. Plusieurs formules équivalentes existent ; la plus claire est reprise ici, issue de `https://fr.wikipedia.org/wiki/Modularité_(réseaux)`.

Ces mesures évaluent la qualité d'une partition ou couverture ; la section suivante s'intéresse à la mesure de similarité entre deux partitions ou couvertures, estimant la similarité entre les résultats des algorithmes de détection de communautés.

### 4.2.2 Mesures de comparaison de deux ensembles de communautés

Pour comparer le recouvrement des communautés détectées par deux algorithmes différents, mais aussi pour mesurer la ressemblance entre deux ensembles d'étiquettes attribuées pour des raisons différentes, la mesure de fiabilité ou **accuracy**, issue de l'apprentissage automatique, est un choix intuitif. Elle mesure le pourcentage d'observations (ici, de nœuds) ayant la même étiquette selon les deux méthodes. Elle fut utilisée, par exemple, pour mesurer la ressemblance entre les communautés issues du graphe DBLP, et le domaine scientifique [Yin et al., 2012].

$$accuracy = \frac{\sum_{u \in \mathrm{U}} \delta(s_u, map(r_u))}{|\mathrm{U}|} \tag{4.10}$$

En équation 4.10, $\delta(x, y)$ vaut 1 si ses arguments sont égaux, 0 sinon. $map(r_u)$ est la meilleure projection [2] des étiquettes $r_u$ dans l'espace des étiquettes de communautés $s_u$.

Dans la même expérience, les auteurs s'intéressent aussi à une mesure plus fine [Yin et al., 2012]. Issue de la théorie de l'information, la mesure **NMI** (*Normalised Mutual Information*) prend en compte la dispersion de chacune des partitions. En équations 4.11 à 4.13, C et C′ sont deux ensembles d'étiquettes (équivalents à nos ensembles de communautés Γ) à comparer. H(C) est l'entropie de C. L'information mutuelle normalisée NMI vaut 1 lorsque les deux ensembles d'étiquettes sont totalement corrélés entre C et C′, indépendamment de la valeur absolue des étiquettes (par exemple, à une permutation près) ; elle vaut par contre 0 s'il n'y a aucune information mutuelle.

$$\mathrm{MI}(\mathrm{C}, \mathrm{C}') = \sum_{c \in \mathrm{C}, c' \in \mathrm{C}'} p(c, c') log \frac{p(c, c')}{p(c)p(c')} \tag{4.11}$$

$$\mathrm{H}(\mathrm{C}) = -\sum_{i=1}^{|\mathrm{C}|} \mathrm{P}(i) \log(p(i)) \tag{4.12}$$

$$\mathrm{NMI}(\mathrm{C}, \mathrm{C}') = \frac{\mathrm{MI}(\mathrm{C}, \mathrm{C}')}{max(\mathrm{H}(\mathrm{C}), \mathrm{H}(\mathrm{C}'))} \tag{4.13}$$

À ces mesures s'ajoutent d'autres classiques, comptant les cas où les nœuds sont étiquetés de manière similaire. Nous reprenons la notation utilisée par [Chen et al., 2014b]. Le nombre de paires de nœuds qui sont dans la même communauté S dans C et dans la même communauté S′ dans C′, est noté $a_{11}$. Pour continuer, $a_{10}$ est le nombre de paires de nœuds dans le même groupe dans C, mais dans des groupes différents dans C′, et similairement pour $a_{01}$. Finalement, $a_{00}$ est le nombre de paires de nœuds qui ne sont dans la même communauté ni selon C, ni selon C′.

L'indice de **Jaccard**, initialement introduit sous le nom de « coefficient de communauté » par [Jaccard, 1901], propose un ratio considérant seulement les paires de nœuds qu'au moins une partition associe. Assez simple à calculer, il est défini en équation 4.14.

$$\mathrm{JI}(\mathrm{C}, \mathrm{C}') = \frac{a_{11}}{a_{11} + a_{10} + a_{01}} \tag{4.14}$$

L'indice de **Rand**, noté *RI* en équation 4.15, dénombre les paires de nœuds « bien classées » selon les deux partitions ; A = $\sum aij$ est le nombre total de paires de nœuds dans le graphe. Cet indice présente l'inconvénient de ne pas suffisamment pénaliser une partition aléatoire, qui n'obtiendrait pas forcément un score de zéro. Aussi, un ajustement est proposé par [Hubert and Arabie, 1985], sous le

2. Dans [Yin et al., 2012], les auteurs se réfèrent à l'algorithme de Kuhn-Munkres, ou « méthode hongroise » dans les problèmes d'affectation [Kuhn, 1955].

nom d'**Adjusted Rand Index** (ARI), défini en équation 4.16 et nécessitant l'introduction d'un terme correctif $\mathrm{M} = \frac{1}{\mathrm{A}}(a_{11} + a_{10}) \times (a_{11} + a_{01})$.

$$\mathrm{RI}(\mathrm{C}, \mathrm{C}') = \frac{a_{11} + a_{00}}{\mathrm{A}} \tag{4.15}$$

$$\mathrm{ARI}(\mathrm{C}, \mathrm{C}') = \frac{a_{11} - \mathrm{M}}{\frac{1}{2}[(a_{11} + a_{10}) + (a_{11} + a_{01})] - \mathrm{M}} \tag{4.16}$$

Ces mesures ne peuvent ni valider une partition, ni donner d'éléments tangibles de la qualité de ses communautés ; elles n'estiment pas la présence d'une chose en commun entre les membres d'un groupe. Cependant, elles permettent de comparer des solutions fournies par différents algorithmes, dégageant par exemple un consensus : certains nœuds peuvent appartenir au même groupe, quelle que soit la méthode choisie.

L'introduction de ces mesures, qu'elles soient topologiques ou de comparaison, permet ainsi d'élaborer et de classer les algorithmes de détection de communautés. L'une des approches suivies par ces méthodes consiste notamment à optimiser la modularité du partitionnement d'un graphe.

## 4.3 Algorithmes de détection de communautés

Cette section s'intéresse au fonctionnement des algorithmes de détection de communautés dans les graphes, mettant de côté les autres types de communautés, non issues de graphes. Dans un premier temps, les algorithmes de partitionnement sont présentés ; ils découpent le graphe en communautés d'intersection nulle. Dans un second temps, des méthodes autorisant la multi-affiliation sont exposées : elles autorisent l'appartenance d'un nœud à plusieurs groupes, simultanément, résultant en une couverture.

### 4.3.1 Partitionnement de graphe

Dans la littérature, il existe de nombreuses propositions de méthodes pour partitionner un graphe. Nous décrivons les algorithmes suivants en détail, car ils couvrent un large spectre d'intuitions et d'approches du problème :

- Girvan-Newman [Newman and Girvan, 2004],
- FastGreedy [Clauset et al., 2004],
- Louvain (parfois aussi nommé Blondel, ou encore Blondel/Louvain) [Blondel et al., 2008],
- InfoMap [Rosvall and Bergstrom, 2007],
- Dominant Flows [Nystuen and Dacey, 1961],
- Markov CLustering for graphs : MCL [Van Dongen, 2000],
- Walktrap [Pons and Latapy, 2005],
- Label Propagation [Raghavan et al., 2007].

Le problème de la détection de communautés est souvent vu comme un problème d'optimisation de la qualité des communautés, dont la mesure principale est la *modularité*, introduite par [Newman and Girvan, 2004]. La définition est rappelée en section 4.2, avec d'autres mesures. Pour le moment, nous nous limitons à son concept : il s'agit du ratio de liens internes aux groupes, par rapport aux connections entre groupes différents.

Comme tout problème d'optimisation, une première approche introduit une méthode gloutonne, itérative : à l'initialisation, chaque nœud est une communauté. Une itération consiste à trouver les deux communautés dont la fusion apporte le plus gros gain marginal de modularité [Newman and Girvan, 2004]. Les itérations successives résultent en un clustering hiérarchique sous la forme d'un

dendrogramme, qui permet de choisir un nombre final de communautés, ou bien en un histogramme des valeurs de modularité dépendant de l'itération, résultant en une partition maximisant la modularité localement : il ne s'agit « que » d'une méthode gloutonne. Cet algorithme est souvent appelé **Girvan-Newman**. Il a une complexité en $O((m+n)n)$ où $n$ est le nombre de nœuds et $m$ le nombre d'arcs dans G. Une première version de cet algorithme se basait sur l'intermédiarité des nœuds, plus longue à calculer [Girvan and Newman, 2002].

Constatant que la plupart des applications recourant à la détection de communautés comportent des graphes peu denses et de faible diamètre, une implémentation astucieuse réduit la complexité moyenne à $O(n \log^2 n)$, sous le nom **FastGreedy** [Clauset et al., 2004]. Les gains sont réalisés d'une part en stockant la matrice des gains marginaux de modularité, et non pas la matrice d'adjacence ; et d'autre part en recourant aux structures de données les plus efficaces pour ce problème. La matrice est stockée sous deux formes, une matrice creuse et un tas ($heap$). Les auteurs revendiquent être les premiers à calculer les communautés sur de gros graphes (400 000 nœuds, 2,4 millions d'arcs, en 2003).

La méthode gloutonne produit souvent de trop grosses communautés, dont le calcul de modularité prend un temps important ; pour profiter de ses atouts en palliant ce défaut, l'algorithme **Louvain** [Blondel et al., 2008] apporte de nouveaux éléments. Une première phase est similaire à la méthode *FastGreedy*, jusqu'à atteindre un minimum local de modularité. Les nœuds sont fusionnés au sein des communautés détectées résultant en un « graphe simplifié », sur lequel une seconde phase, semblable à la première, est appliquée. Ce processus est répété plusieurs fois, jusqu'à trouver un maximum final. Chaque nouvelle phase est mécaniquement plus rapide que la précédente, puisque le graphe simplifié présente nettement moins d'arcs et de nœuds. La complexité empirique sur des réseaux grands, mais peu denses [3], est en $O(n \log n)$.

La complexité des réseaux, sociaux ou biologiques, a inspiré le besoin d'avoir une carte pour y naviguer plus aisément, un équilibre pour représenter une réalité simplifiée du graphe. Suivant cette intuition, la méthode **InfoMap** [Rosvall and Bergstrom, 2007] représente le flot d'information dans un graphe : une marche aléatoire parcourt le graphe, attribuant des noms aux nœuds par la méthode de codage de Huffman [4]. Il s'agit de codes binaires, changeant très légèrement d'un nœud vers le suivant. Une seconde phase consiste à compresser ces noms de nœuds, les groupant en « modules » qui correspondent aux communautés recherchées.

Issu de la recherche en géographie, la méthode des **flots dominants**, ou *Dominant Flows* [Nystuen and Dacey, 1961], vise à regrouper ensemble des villes pour former des bassins économiques. Les échanges entre villes sont quantifiés ; un score d'importance correspond à la somme des flux entrants. Afin de prendre en compte les flots indirects (de $a$ vers $b$ puis $c$), la matrice d'adjacence A est normalisée puis remplacée par $B = \sum_i A^i$, c'est-à-dire la somme des flots de longueur 1, puis de longueur 2, *etc*. Sur cette « matrice d'influence indirecte », les villes/nœuds sont rattachées à la cité vers laquelle ils envoient leur plus grand flux ; par définition, les villes les plus importantes ont leur flot sortant vers des villes plus petites. Souvent exploitée en géographie sur des graphes relativement petits, les *Dominant Flows* sont réputés sensibles au bruit : l'ajout ou le retrait d'arcs résulte en une partition très différente [Queyroi et al., 2015].

La matrice d'adjacence d'un graphe, où $A_{i,j}$ vaut 1 s'il y a un arc $e_{i,j}$ et 0 sinon, est facile à convertir en une matrice des probabilités $\tilde{A}$ de passer d'un nœud à un autre. Une telle notion permet de réaliser des marches aléatoires (un marcheur se déplace dans le graphe, choisissant sa prochaine destination par tirage au sort). L'algorithme **Markov Clustering** MCL, par une approche totalement algébrique, alterne une phase d'expansion, c'est-à-dire de réalisation de la marche aléatoire par multiplication par $\tilde{A}$, à une phase d'inflation, reposant sur la normalisation des colonnes, qui renforce les liens internes aux groupes. La répétition de ces deux phases produit rapidement des communautés

3. Site de l'auteur : `https://perso.uclouvain.be/vincent.blondel/research/louvain.html`
4. Méthode de compression : `https://fr.wikipedia.org/wiki/Codage_de_Huffman`

denses [Van Dongen, 2000]. Une version naïve de la méthode a une complexité cubique ; cependant le calcul matriciel est aisé à paralléliser. En posant $k$ le nombre maximal autorisé de connexions sortantes d'un nœud (hypothèse que le graphe est peu dense), la complexité est en $O(nk^2)$ selon son créateur [5].

Exploitant des marches aléatoires, l'algorithme **Walktrap** profite des forces du clustering hiérarchique : en partant de chacun des nœuds, une marche aléatoire permet de définir une distance moyenne entre chaque paire de nœuds ; dans une seconde étape, l'algorithme construit le dendrogramme représentant le graphe, à partir duquel il est aisé de tirer le nombre de communautés voulues [Pons and Latapy, 2005]. La complexité est en $O(mn^2)$ dans le pire des cas. Pour un « graphe réel » (peu dense, petit monde, invariant d'échelle), elle est en $O(n^2 \log n)$.

Sans fonction à optimiser, l'algorithme de **Label Propagation** [Raghavan et al., 2007] suggère une contamination des nœuds par les étiquettes de communautés. À l'initialisation, chaque nœud reçoit un label différent. Le cœur de la méthode consiste à attribuer itérativement à chaque nœud, l'étiquette la plus présente dans son voisinage direct. En cas d'égalité, un tirage au sort aléatoire uniforme est réalisé. La convergence est rapide dans les sous-graphes denses et isolés ; cependant un phénomène d'oscillation peut apparaître, notamment lorsque le graphe est bipartite : chacune des deux parties récupère à chaque itération l'étiquette de l'autre, et la méthode ne trouve alors pas d'équilibre. L'algorithme s'arrête lorsque chaque nœud du graphe affiche la même étiquette que la majorité de ses voisins. Chaque itération a une complexité en $O(m)$, et les auteurs mentionne une « convergence significative après 5 itérations », sans prouver la complexité moyenne. Empiriquement, les temps de calcul sont bons grâce notamment à une parallélisation aisée.

Une comparaison de ces méthodes sur de grands graphes (environ 400 000 nœuds) distingue *Louvain*, qui est le plus rapide pour des qualités comparables [Cazabet and Amblard, 2011] : l'expérience inclut *CFinder* (recherche et fusion de cliques) [Adamcsek et al., 2006], *FastGreedy*, *iLCD* (méthode multi-agents introduite par les auteurs), *InfoMap* et *Louvain*. Les graphes utilisés sont générés artificiellement, couvrant un grand intervalle de nombre de nœuds, de 0 à 400 000. La figure 4.1 affiche les temps de calcul (sur un PC simple, 4Go de RAM) pour ces méthodes en fonction du nombre de nœuds du graphe ; ni *FastGreedy* ni *Cfinder* n'arrivent à tenir en mémoire au-delà de 200 000 nœuds. *Louvain* est le plus rapide dans cette expérience.

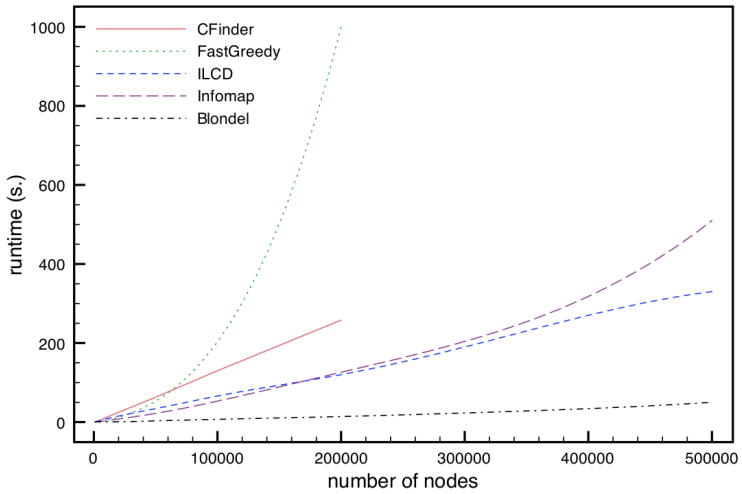



FIGURE 4.1 – Comparaison des temps de calcul par rapport à la taille du graphe, issu de [Cazabet and Amblard, 2011]. NB : Louvain apparaît sous le nom Blondel.

Une autre comparaison se penche sur les mentions d'États (environ 200 sur notre planète) dans

5. Site web dédié à MCL : `https://micans.org/mcl/`

des articles de presse [Queyroi et al., 2015]. La co-occurrence de deux nœuds-nations ajoute un arc entre elles, résultant en un graphe. L'objectif est de vérifier la validité thématique et géographique des communautés. Dans cette expérience sont comparés *Louvain*, *Label propagation*, *Dominant Flows* et *Markov CLustering* : une distance est calculée entre les partitions obtenues. À première vue, *Louvain* est très éloigné des autres algorithmes, qui proposent des groupes plus nombreux, plus petits, dont l'agrégation permet de retrouver une partition similaire à celle de *Louvain*, maximisant la modularité.

### 4.3.2 Couvertures : appartenance multiple à des communautés

La recherche de groupes très fortement connectés débouche naturellement sur un extrême, la notion de *clique* : un ensemble de nœuds tous reliés les uns aux autres. Une $k-$clique est une clique contenant $k$ nœuds. La condition de forte liaison interne est alors satisfaite. Afin de relâcher cette contrainte, une méthode de percolation fusionne les cliques qui partagent tous leurs nœuds sauf un : chaque membre reste lié à au moins $k$ autres membres, et la clique finale est plus grande. Certains nœuds, présents dans deux cliques différentes, sont parties prenantes de deux communautés. Le choix de la valeur de $k$ permet d'orienter la recherche vers des communautés suffisamment grandes (et donc connectées). Cette approche par cliques présente des résultats prometteurs sur des réseaux de protéines [Palla et al., 2005]. Une implémentation sous le nom de **CFinder** [Adamcsek et al., 2006] résulte en une complexité exponentielle (pire des cas), mais qui se concrétise en des temps réalistes lorsque le graphe est peu dense.

Dérivé des cliques, le concept de **k-cores** permet d'obtenir des communautés avec recouvrement. Un $k-core$ est le plus grand ensemble connexe de nœuds, chacun étant connecté à au moins $k$ autres membres [Giatsidis et al., 2013]. Pour des valeurs de $k$ assez faibles, et selon la structure du graphe, un nœud peut appartenir à deux communautés (par exemple, s'il agit comme un pont entre deux groupes, en étant fortement connecté aux deux groupes). La détection des $k-cores$ n'est pas rapide, et nécessite un réseau assez segmenté : un $k-core$ contenant la majorité du graphe n'est pas très utile ni exploitable.

L'algorithme **CONGA** (*Cluster-Overlap Newman-Girvan Algorithm*) [Gregory, 2007] se base sur la mesure d'intermédiarité pour extraire des couvertures. L'intermédiarité d'un nœud $i$ est le nombre de plus courts chemins entre toutes les paires de nœuds du graphe, passant par $i$ : les nœuds de forte intermédiarité sont les points de passage favoris entre parties distantes du graphe, c'est-à-dire entre communautés différentes. À chaque itération, une décision doit être prise : soit le nœud d'intermédiarité maximale est coupé en deux, chaque moitié correspondant à une communauté, désormais isolée de l'autre ; soit l'arc d'intermédiarité maximale est retiré, simplifiant la structure du graphe pour finir par obtenir un dendrogramme (clustering hiérarchique). Au cours du processus, un nœud peut être coupé plusieurs fois : il appartiendra alors à plusieurs groupes. Cette méthode est coûteuse en temps, avec dans le pire des cas une complexité en $O(m^3)$ : selon le graphe, le fait de couper un nœud en deux peut augmenter ou diminuer la complexité du calcul restant.

Une version plus rapide, **CONGO** (le dernier O signifie *optimisé*) [Gregory, 2008], se base sur l'intermédiarité *locale* : il s'agit du nombre de plus courts chemins de longueur maximale $h$, où $h$ est un paramètre de l'algorithme (2 ou 3 sont des valeurs recommandées pour limiter les calculs). Ainsi, pour calculer l'intermédiarité de $i$, il n'y a pas besoin de parcourir l'ensemble des paires de nœuds du graphe, puisque *CONGO* se limite aux nœuds proches de $i$. Une complexité en $O(n \log n)$ est revendiquée pour des réseaux épars.

Sur un autre paradigme, les auteurs de *CONGO* ont plus tard introduit l'algorithme **COPRA** (*Community Overlap PRopagation Algorithm*) [Gregory, 2010], qui se base sur le partitionnement par l'algorithme de *Label Propagation* [Raghavan et al., 2007], en y ajoutant un paramètre $v$ : le nombre maximal de communautés auxquelles un nœud peut appartenir. Durant le processus, chaque nœud peut ainsi recevoir jusqu'à $v$ étiquettes de voisins, associées à un poids d'appartenance ; en fin d'itération, les étiquettes de poids trop faible sont retirées afin de ne pas propager du bruit. Le fonc-

tionnement global est très proche de *Label Propagation*, résultant en une faible complexité moyenne pour les graphes épars, en $O(v^3 n)$. Les valeurs recommandées pour $v$ dépendent du problème, mais sont susceptibles d'être inférieures à 5 pour la plupart des applications.

S'inspirant aussi de la propagation des étiquettes, **SpeakEasy** [Gaiteri et al., 2015] combine une approche locale, où les nœuds adoptent les étiquettes de leurs voisins, avec une visibilité globale : les étiquettes sont (dé)favorisées selon leur popularité, leur fréquence dans le graphe. La plupart des nœuds finissent avec une unique étiquette, à l'exception des ponts entre communautés, qui sont multi-affiliés. Les auteurs donnent accès à leur code ainsi qu'à des exemples sur leur site [6].

La méthode **SLPA** (*Speaker-listener Label Propagation Algorithm*) [Xie and Szymanski, 2012] ajoute une sélection aléatoire dans la méthode *Label Propagation* : à chaque étape, un nœud est sélectionné pour être le *listener* (auditeur), et chacun de ses voisins devient *speaker*, choisissant aléatoirement l'un des couples (étiquettes, poids) à retransmettre à son voisinage. L'auditeur ajoute l'étiquette la plus populaire parmi celles qu'il reçoit : il peut conserver ainsi un grand nombre d'étiquettes, correspondant à différentes communautés. L'algorithme nécessite deux paramètres : $r$ permet d'éliminer, après la dernière itération, les couples (étiquettes, poids) de poids trop faible ; T est le nombre maximal d'itérations, choisi par l'utilisateur. En conséquence, la complexité est en $O(m\mathrm{T})$ ; les auteurs proposent une comparaison avec d'autres méthodes dans un second article [Xie et al., 2013].

Une approche locale, **OSLOM** (*Order Statistics Local Optimization Method*) [Lancichinetti et al., 2011], explore le graphe à la recherche de clusters « significatifs » : des groupes denses de nœuds. Réalisée à plusieurs niveaux d'échelle, cette démarche permet de trouver des groupes de tailles différentes, qui peuvent éventuellement se recouvrir. Ainsi, le « problème de résolution » de la modularité est évité (l'optimisation de la modularité tend à retourner des communautés de taille comparable, éliminant les groupes trop petits ou trop gros pour un graphe donné). Deux phases de nettoyage fusionnent ou éliminent les communautés superflues ou se recouvrant en totalité ; finalement les communautés elles-mêmes sont regroupées, facilitant la lecture du réseau. La complexité théorique n'est pas connue ; les mesures effectuées sur des graphes artificiels suggèrent qu'elle est linéaire en $n$.

La détection de couvertures est souvent effectuée soit par des algorithmes de complexité prohibitive (détection de cliques notamment), soit par des méthodes (presque) linéaires, mais difficiles à valider. Les poids d'appartenance à un groupe ajoutent une subtilité à prendre en compte : un groupe est-il réel si aucun de ses membres n'y appartient à plus de 20% ? Cependant, ce point de vue semble plus proche de la réalité que l'attribution simple calculée par les méthodes de partitionnement.

### 4.3.3 Synthèse des algorithmes de détection de communautés

Dans cette section, deux manières d'obtenir des communautés ont été présentées : les partitions et les couvertures, selon le nombre de groupes auxquels un nœud peut appartenir. Dans les deux cas, quelques approches et intuitions rassemblent les propositions d'algorithmes : l'optimisation d'une mesure de qualité des groupes, les marches aléatoires dans un graphe, la propagation d'étiquettes, et la détection de structures particulières, notamment les cliques. Ces approches sont reprises en table 4.1, selon qu'il s'agisse d'obtention de partition ou de couverture.

S'il semble légitime d'autoriser un nœud à appartenir à plusieurs groupes, cela reste un point complexe à calculer et difficile à valider. Aussi nous préférons nous focaliser sur les méthodes de partition, parmi lesquelles *Louvain* présente l'une des meilleures complexités, et retourne des partitions de bonne modularité.

6. `http://www.cs.rpi.edu/~szymansk/SpeakEasy/`

TABLEAU 4.1 – Récapitulatif des algorithmes présentés

| | Approche | Algorithme | Complexité | Référence |
|---|---|---|---|---|
| Partition | Optimisation (ex : de Q) | Girvan-Newman | $O((m+n)n)$ | [Newman and Girvan, 2004] |
| | | FastGreedy | $O(n\log^2 n)$ | [Clauset et al., 2004] |
| | | Louvain | $O(n\log n)$ | [Blondel et al., 2008] |
| | Marches aléatoires | Dominant Flows | - | [Nystuen and Dacey, 1961] |
| | | MCL | $O(nk^2)$ | [Van Dongen, 2000] |
| | | InfoMap | - | [Rosvall and Bergstrom, 2007] |
| | | WalkTrap | $O(n^2\log n)$ | [Pons and Latapy, 2005] |
| | Propagation d'étiquettes | Label Propagation | $O(m)$ | [Raghavan et al., 2007] |
| Couverture | Optimisation (ex : de Q) | CONGO | $O(n\log n)$ | [Gregory, 2008] |
| | | OSLOM | $O(n)$ | [Lancichinetti et al., 2011] |
| | Propagation d'étiquettes | COPRA | $O(v^3 n)$ | [Gregory, 2010] |
| | | SpeakEasy | - | [Gaiteri et al., 2015] |
| | | SLPA | $O(m\mathrm{T})$ | [Xie and Szymanski, 2012] |
| | Cliques | CFinder | $O(e^n)$ ? | [Adamcsek et al., 2006] |
| | | k-cores | $O(e^n)$ ? | [Giatsidis et al., 2013] |

## 4.4 Applications à des réseaux sociaux numériques

Cette section présente des exemples de détection de communautés appliquée aux réseaux sociaux numériques. Dans un premier temps, des travaux se penchent sur l'exploitation d'une seule information : un graphe représentant l'activité sur une plate-forme. Dans un second temps sont exposées des approches « hybrides », qui adjoignent à ce graphe un autre type d'information, par exemple le texte de pages Web. Enfin des exemples d'exploitation des communautés sont présentés, dont la détection d'événements.

### 4.4.1 Approches classiques

Sur Twitter, les utilisateurs peuvent créer des « listes » d'abonnements, regroupant sur une même page les contenus produits par d'autres. Cette fonctionnalité a une vocation thématique, donnant à chacun les clés pour trier leur flux entrant de tweets. Pour trouver des comptes actifs sur une même thématique, un graphe est construit, dont les nœuds sont des *listes*. Lorsque deux listes sont suffisamment similaires (selon une mesure dérivée de *NMI*), un arc est placé entre elles. L'algorithme *OSLOM* détecte ensuite, sur ce graphe, des communautés (avec superposition possible) de listes de comptes ; ces listes étant nommées, des noms sont proposés pour étiqueter les clusters [Greene et al., 2012].

La disponibilité des données issues de Twitter, notamment via la publication [7] d'un grand graphe [Kwak et al., 2010] a suscité un engouement pour le graphe des abonnements. Depuis, il a été observé que cette relation ne représente pas bien l'interaction réelle entre les comptes [Lim and Datta, 2016]. Des éléments tangibles de comparaison sont proposés entre partitions du graphe des abonnements ou des « intérêts partagés », comparant le coefficient de clustering, le degré moyen et la longueur moyenne des plus courts chemins au sein d'une communauté.

Un argument supplémentaire contre le graphe des abonnements consiste en son inertie, par opposition aux retweets qui sont, par nature, viraux. Autour d'événements particuliers, l'exploitation des fonctionnalités de retweet et de mention permet de relier les utilisateurs actifs entre eux. Dans un second temps, un partitionnement par marches aléatoires fournit des communautés, dont les étiquettes sont calculées à partir des textes des tweets émis par leurs membres [Tyshchuk et al., 2014]. Cette démarche clarifie l'information provenant des médias sociaux, puisqu'elle identifie le contenu, ainsi que les groupes sociaux émergents qui le propagent ; les auteurs l'appliquent à des tweets émis lors de combats en Syrie, en 2013.

7. Par Stanford, sur `https://snap.stanford.edu/data/twitter7.html`, désormais retiré à la demande de Twitter.

### 4.4.2 Approches hybrides

Les approches hybrides sont caractérisées par l'entrelacement entre liens d'interaction et liens sémantiques. Les premiers travaux exploitaient le référencement hypertextuel [Cohn and Hofmann, 2001] : chaque page Web est source d'arcs dirigés vers d'autres pages. Dans un souci d'indexation du Web (le graphe formé par les liens d'une page à l'autre), les auteurs sont à la recherche d'une part de communautés thématiques, et d'autre part d'autorités sur une thématique. Pour faciliter cette recherche, les arcs sont pondérés par une mesure de similarité calculée à partir des textes des pages.

Afin de détecter des communautés thématiques d'utilisateurs sur des forums, une pondération par la similarité sémantique est appliquée sur les arcs reliant les utilisateurs. Le poids des arcs provient du contenu, mais leur existence dépend de l'interaction, par exemple le fait qu'un utilisateur réponde à un billet de blog émis par un autre [Liu et al., 2009]. Les auteurs proposent une validation via une expérience supplémentaire : la prédiction de lien manquant dans le graphe. Issue de l'analyse textuelle, la modélisation des thématiques d'intérêt pour chaque compte est réalisée par une technique dérivée de LDA (*Latent Dirichlet Allocation*) [Blei et al., 2003].

L'approche inverse consiste à améliorer le modèle des thématiques par l'exploitation du réseau entre auteurs [Nallapati et al., 2008]. Sur le réseau de citations scientifiques, un LDA regroupe ensemble des documents sur la base des fréquences des mots ; la probabilité que deux documents soient de même thématique est d'autant plus grande s'il y a un lien de citation entre eux.

Sans explicitement construire de graphe, un modèle probabiliste bayésien fait le pont entre les notions de communautés et de thématiques. Les e-mails échangés dans une compagnie, et notamment le corpus *Enron* (environ 100 000 e-mails, publiés suite à la fermeture de l'entreprise en 2001), fournissent un support pertinent pour relier les personnes, les communautés, les mots et les thématiques des échanges : un réseau bayésien représentant chacune de ces relations par des probabilités conditionnelles détermine à la fois les T thématiques et les C communautés (nombres choisis *a priori*) [Zhou et al., 2006]. Dans ce modèle, le graphe est sous-entendu par le réseau bayésien : les fréquences de contact entre personnes sont régies par la probabilité qu'elles soient dans la même communauté.

Enfin, la combinaison des deux types d'information (du texte et des relations d'interaction) peut être réalisée tard, pour produire deux vues complémentaires du réseau [Yin et al., 2012]. La modélisation des thématiques des textes d'une part, et la détection des groupes à partir de caractéristiques topologiques dans le graphe des abonnements d'autre part, respectent les contraintes suivantes : une communauté peut être active sur plusieurs thématiques, et un même thème peut être présent dans plusieurs groupes. La validité des communautés ainsi obtenues est calculée par la fiabilité et l'information mutuelle normalisée (NMI) entre le découpage par thématiques, et celui par communautés. Cette méthode est appliquée sur deux corpus, l'un issu de DBLP, l'autre constitué de tweets.

### 4.4.3 Exemples d'utilisation : détection d'événement, de thématique, d'oppositions

Les communautés sont un bon moyen de confirmer ou d'infirmer un signal détecté par ailleurs : par exemple, un module de détection d'événements, purement basé sur les dépassements de seuils de fréquences de mots, risque de générer de nombreux faux positifs. En associant les « explosions » de mots-clés (*burstiness*) avec leurs auteurs, mis en relation dans le graphe des abonnements, le système *MABED* (*Mention Anomaly Based Event Detection*) est capable de réduire le bruit et donc, de mieux détecter des événements [Guille and Favre, 2015]. Plus précisément, la méthode *Louvain* est utilisée : le fait que plusieurs membres de la même communauté évoquent ce qui pourrait constituer un événement renforce le signal. L'exemple proposé contient environ 50 000 utilisateurs reliés par plus de 5 000 000 d'abonnements.

Les algorithmes de détection de communautés sont employés dans plusieurs applications touchant directement aux réseaux sociaux numériques, que ce soit sur des graphes d'utilisateurs, ou de termes employés. Les communautés obtenues sont susceptibles d'évoluer au fil du temps, et il est difficile de

garder la trace d'un groupe au gré des ajouts ou retraits de membres, ainsi que des fusions ou scissions de groupes.

Réalisée sur un graphe de citations extrait d'*Arxiv*, une expérience compare la stabilité des résultats de *Louvain*, *FastGreedy* et *Walktrap* ; dans un premier temps en déroulant le jeu de données, étalé sur plusieurs années ; dans un second temps en retirant aléatoirement des nœuds et arcs du réseau. La stabilité des groupes détectés face au retrait d'un nœud sert de mesure de la qualité de l'algorithme [Aynaud and Guillaume, 2010]. Ces travaux ouvrent la voie à un suivi des communautés dans le temps : la détection n'est qu'une première étape.

Pour évaluer l'importance des camps en présence sur un débat, un graphe contenant les réponses (*replies*) entre utilisateurs mentionnant le même hashtag #IPCC (*Intergovernmental Panel on Climate Change*, qui publie régulièrement des rapports sur le changement climatique) fournit des communautés concordant, selon des annotateurs humains, avec des soutiens et détracteurs de la réalité du changement climatique [Pearce et al., 2014] : en cela, les groupes détectés semblent pertinents. Cependant, ni leur force d'interaction interne des groupes ni leur évolution ne sont étudiées. L'utilisation du logiciel Gephi [8] fournit les illustrations du débat public, qui a eu lieu sur Twitter, entre climatosceptiques et écologistes.

Dans ces travaux, la description de l'activité sur un segment de réseau social passe par son découpage en communautés. Nous avons vu que des approches hybrides, ajoutant des informations provenant du texte sur un graphe de relations, complète cette description qui prend alors la forme d'un graphe enrichi, ou d'une modélisation bayésienne.

L'analyse des réseaux sociaux par le prisme des graphes et des communautés est particulièrement efficace, notamment pour des applications spécifiques de détection d'événement, ou encore pour quantifier les oppositions sur un débat politique. Cependant, la simple détection de groupes ne suffit pas ; cette lecture par groupes doit être complétée par l'étude des contenus et des comptes actifs.

## 4.5 Synthèse de la détection et analyse de communautés

La dimension sociale de l'activité sur les réseaux sociaux numériques, reliant directement les comptes utilisateurs, génère de la donnée qu'il est intuitif de modéliser sous forme de graphes. L'émergence, dans ces graphes, de structures denses, c'est-à-dire de communautés d'utilisateurs, ajoute une nouvelle dimension d'analyse. Bien qu'elles ne soient pas rigoureusement définies, les communautés thématiques, groupes d'utilisateurs actifs sur un même thème, constituent des éléments de quantification des forces en présence autour d'un débat, ou de confirmation de l'identification d'un événement.

L'approche retenant notre attention consiste à identifier des communautés à partir de graphes ; les outils mathématiques existants sont déjà nombreux pour cette étape. Marches aléatoires, optimisation de la modularité et propagation d'étiquettes sont les principaux paradigmes des algorithmes de détection de communautés. Cependant, certains aspects sont parfois absents, ne considérant pas la direction ou les poids des arcs, par exemple.

La complexité des algorithmes proposés pose également parfois problème : une détection de cliques en temps exponentiel n'est pas réaliste sur un graphe avec plusieurs milliers de nœuds. Pour gagner en complexité, *Label Propagation* mise sur l'aléatoire, au risque de ne pas converger : dans tous les cas, le résultat renvoyé par l'algorithme doit être évalué.

Cette critique bénéficie d'une variété de mesures topologiques, évaluant différents aspects de la force de liaison interne des communautés, par exemple par leur densité et leur modularité. De même, les indices de Rand et de Jaccard aident à mesurer la similarité de deux ensembles de communautés : s'ils sont issus de deux méthodes différentes, ces mesures quantifient l'accord entre les méthodes. Un consensus regroupe les communautés identifiées par plusieurs méthodes différentes, à quelques

8. Libre, il permet d'analyser et visualiser des graphes. `https://gephi.org/`

nœuds près.

Dans les travaux existants, portant sur la détection ou l'analyse des communautés présentes sur les réseaux sociaux numériques, et malgré une présence fréquente de la notion de « communauté thématique », le texte des messages échangés est peu voire pas pris en compte. La thématique s'identifie parfois par un hashtag ou une célébrité. Le focus est alors mis soit sur la détection de communautés dans un graphe, sans considérer le texte [Aynaud and Guillaume, 2010, Lim and Datta, 2012], soit sur le texte, la similarité sémantique seule servant à délimiter les groupes d'utilisateurs [Rosa et al., 2011, Amelio and Pizzuti, 2015].

Lorsqu'ils contiennent un graphe, de nombreux travaux n'exploitent pas les différents types de relations, se contentant de construire un graphe sur la base d'une intuition (par exemple, le graphe des abonnements), sans discuter du sens ou de la pertinence, ni considérer d'autres relations. Or les nœuds peuvent être reliés à la fois par un abonnement qui dure, par des retweets ponctuels et par une discussion sous la forme de réponses ; des relations et interactions aux significations parfois différentes.

Finalement, la détection de communautés, en soi, n'est pas l'objectif final : il faut extraire du sens des résultats. Nous pensons que cela passe par une caractérisation plus profonde des communautés, par plusieurs pistes. Une première voie consiste à combiner les thématiques de chacun pour calculer des scores de cohésion et d'influence, au niveau du groupe : ainsi l'action coordonnée d'un groupe de petits comptes serait identifiée comme un objet *influenceur*, qualifiant le groupe d'acteur-clé. Une seconde voie repose sur l'analyse des types de comptes et de comportements adoptés au sein d'un groupe, menant à l'identification des meneurs et suiveurs : l'étude de la dynamique de groupe devrait en sortir renforcée.

Dans le chapitre 7, nous proposerons un modèle pour construire les graphes représentant l'activité du réseau social, et nous introduirons des mesures pour caractériser les communautés qui y apparaissent, en prenant en compte les textes des messages publiés par leurs membres. Cette analyse textuelle, visant à déterminer l'opinion, c'est-à-dire la thématique et la polarité des messages, constitue notre première contribution. Elle est exposée en chapitre 5.

# Deuxième partie

# Contributions théoriques

# CHAPITRE 5

## La contextonymie pour détecter la posture dans le tweet

Les chapitres précédents ont exposé un ensemble de méthodes permettant de mieux comprendre le fonctionnement des réseaux sociaux numériques. Notamment, la première difficulté concerne l'exploitation de la masse importante de texte à disposition sur ces plateformes.

Dans ce chapitre, nous nous concentrons sur une tâche, la détection de posture, pour mettre en avant la difficulté provenant du texte même : les spécificités du « langage social », ou *sociolecte*, sont un frein aux systèmes nécessitant des ressources linguistiques. Pour lever ce verrou, nous utilisons et adaptons aux tweets une approche peu connue, la contextonymie.

Ce chapitre est structuré comme suit : la section 5.2 formalise la tâche à accomplir, la détection de posture, avec des exemples mettant en exergue les difficultés rencontrées, ainsi que le verrou scientifique identifié. Une revue de l'état de l'art de la détection de posture a été proposée en chapitre 2, et plus particulièrement la section 2.5.2. Une proposition de méthodes « simples » permet d'avoir une première solution, et de mieux comprendre la difficulté tant dans la détection de posture que dans le traitement des textes issus des réseaux sociaux.

Pour adresser le défi des sociolectes, la section 5.3 expose la méthode de construction de la ressource linguistique de contextonymie : invention, histoire, ressources nécessaires et algorithme sont décrits en détails. Ensuite, la section 5.4 illustre les contextonymes et contextosets, permettant au lecteur d'en percevoir tout l'intérêt. Enfin, l'utilisation des contextosets pour améliorer les classifieurs de posture proposés préalablement donne lieu à des résultats chiffrés en section 5.5, puis à une section de discussion.

## 5.1 Détection de posture dans des tweets

Parmi la variété des tâches envisageables sur le traitement de textes émis sur des réseaux sociaux, nous avons identifié, dans le chapitre 2, la détection de posture. En effet, Twitter est communément utilisé pour exprimer des avis, vues et ressentis sur des thématiques variées, allant des revues de produits aux débats politiques. Dans ce domaine, la détection de posture permet de résumer un tweet au couple (cible, posture). Par exemple, dans le thème de la campagne électorale américaine de 2016, une *cible* pourrait être Hillary Clinton, ou Donald Trump, et la posture serait un soutien ou une opposition vis-à-vis d'un candidat. Cette tâche consiste donc à déterminer, à partir du texte d'un message, s'il exprime un soutien, une opposition ou l'absence d'avis vis-à-vis d'une entité, la cible.

### 5.1.1 Description du problème

Les tweets sont des messages courts, qui contiennent des orthographes inventives dont le sens est souvent implicite. Ils diffèrent cependant des SMS [Gotti et al., 2013] : les tweets sont des messages publics, de grande diffusion, sans nécessairement de destinataire. Les SMS sont des messages strictement privés, de point à point. De par son côté social, Twitter favorise la diffusion de mots inventés et de fautes de frappe intentionnelles [Maynard et al., 2012]. En outre, les tweets contiennent des termes spécifiques tels que les mots-dièse (hashtags) et les mentions d'utilisateurs. Les hashtags sont des mots ou phrases précédés par un dièse #, dont le sens est souvent maintenu dans le tweet. Par exemple, « #voteforyou » peut remplacer « vote for you ». En revanche, ils ne sont pas systématiquement décomposables ou compréhensibles : #MLP peut se référer à une femme politique française, Marine Le Pen, ou à un dessin animé pour jeunes enfants, My Little Pony. Leur utilisation et durée de vie sont très variables : des émissions de télévision ou des partis politiques se maintiennent dans le temps, tandis que certains hashtags sont éphémères. Enfin, tant les hashtags que les mentions peuvent aussi apparaître dans les tweets sans aucun rôle syntaxique, uniquement comme des labels ou des tentatives d'attraction. Ces éléments caractéristiques aux tweets ont un impact sur leur traitement automatique.

Des techniques d'extraction d'opinion fonctionnent bien sur des textes, surtout lorsqu'ils sont rédigés en anglais correct [Bird et al., 2009]. La difficulté est ici de transposer ces méthodes sur des tweets. Cependant, la plupart des algorithmes orientés tweets considèrent les mots comme des atomes, sans prendre en compte la relation avec leur contexte ; cela génère de l'ambiguïté, car la plupart des mots permettent plus d'une interprétation en fonction du contexte d'utilisation.

Pour pallier ce manque, nous proposons de recourir aux *contextonymes*, suivant l'intuition que l'environnement d'un mot, c'est-à-dire les quelques termes précédents et suivants, peut permettre d'affiner l'interprétation qu'on en donne, et ainsi rendre possible la compréhension des mots ambigus ou jusqu'alors inconnus. Deux mots sont contextonymes s'il sont fréquemment utilisés dans un même contexte ; un ensemble de mots contextonymes entre eux est appelé *contextoset*.

### 5.1.2 Approche proposée pour la détection de posture

Dans ce chapitre, nous nous intéressons à la tâche de détection de posture dans le tweet. Deux approches classiques existent : d'une part, l'approche par analyse de sentiment, à base de dictionnaires ; d'autre part, l'approche par classification sur un corpus annoté (apprentissage supervisé). À partir de ces deux approches, nous utilisons deux méthodes qui servent de *baselines*.

Parallèlement, nous procédons à la construction d'une ressource linguistique, un dictionnaire de contextosets, appris sur *GenTweets*, un corpus de plusieurs millions de tweets. Nous exploitons cette ressource pour améliorer les deux approches précédentes. La tâche de *stance detection* (détection de posture) proposée par SemEval en 2016 fournit un corpus de test et des compétiteurs, ce qui permet par la suite de situer notre contribution au regard de l'état de l'art.

## 5.2 Les difficultés des dictionnaires de sentiment pour traiter les tweets

Dans cette section, nous formalisons le problème de la détection de posture. Puis nous mettons en avant la difficulté rencontrée par les approches classiques, notamment à base de dictionnaires de sentiment, sur les tweets.

### 5.2.1 Définitions pour la détection de posture dans un tweet

Introduite en définition 5.2.1 et suivantes, la détection de posture consiste à analyser un **texte** vis-à-vis d'une **cible**, pour en déterminer la **posture** (*stance* en anglais). La figure 5.1 montre les entrées et la sortie possible d'un tel analyseur. Par intuition, chaque cible peut être associée à un vocabulaire

spécifique, ce qui nécessiterait aussi un analyseur spécifique, au moins dans sa mise au point. Ce module peut être basé sur de l'apprentissage, des règles ou autre.

**Definition 5.2.1.** *Texte :* la chaîne de caractères d'entrée. Le format court du tweet est propice à ne contenir qu'une seule phrase, que nous supposons porteuse d'une seule Posture envers une cible donnée.

**Definition 5.2.2.** *Cible :* l'entité envers laquelle une posture doit être déterminée, choisie au préalable. Celle-ci n'est pas toujours mentionnée explicitement dans le Texte.

**Definition 5.2.3.** *Posture :* la polarité du message envers la cible de l'analyse. Elle est représentée par un label, et peut être *Positive*, *Négative* ou *Neutre*.

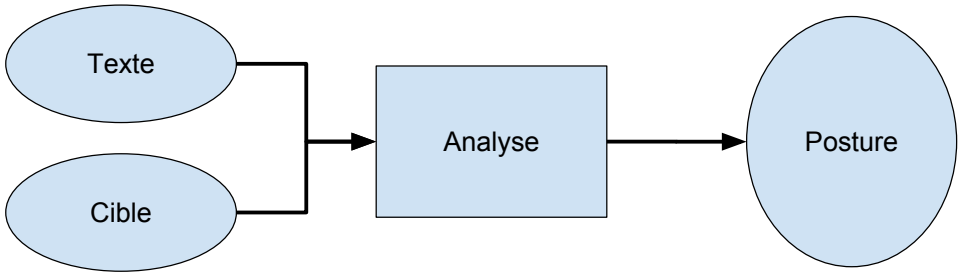



FIGURE 5.1 – Vue générale de la détection de posture

Nous donnons ci-dessous un exemple de tweet, de posture positive envers la cible, qui est Hillary Clinton. Les cinq premiers mots suffisent pour déterminer que le message soutient la candidate ; la posture se trouve renforcée par le hashtag de la campagne.

| Hillary is our best choice if we truly want to continue being a progressive nation. #Hillary2016 #Ohio |
|---|

Cependant, le problème n'est pas toujours aussi aisé. Ciblant le changement climatique (« Climate change is a real concern »), le tweet suivant, par son côté ironique, discrédite les alarmistes - et tous ceux qui voient des signes du réchauffement climatique - en y associant l'été, et sa hausse normale des températures : la posture est climato-sceptique.

| Global Warming ! Global Warming ! Global Warming ! Oh wait, it's summer. |
|---|

Enfin, les messages peuvent aussi être hors sujet, en ne faisant aucune référence à la cible ; ou bien concerner la cible sans contenir d'élément permettant de déduire la posture, et être neutres ou objectifs. Dans tous ces cas de figure, nous utilisons le même label, « NEUTRE » : nous considérons que la posture n'est pas exprimée dans la phrase. Par exemple, le message suivant, présent dans un corpus d'entraînement à la détection de posture [1] au regard de la cible « Atheism », ne permet pas de conclure d'une quelconque posture : aucune référence à la religion n'est présente.

| Alot of angry people in this world. Peace to all. #love |
|---|

Pour ce dernier tweet, notons que l'ajout d'un hashtag peut clarifier cet état de fait : « #Jesus » en fin de tweet permet de rattacher tout l'énoncé aux principes chrétiens. À l'opposé, un hashtag supplémentaire tel que « #NoHolyWar » (pas de guerre sainte), serait une critique de la religion, promouvant l'athéisme.

1. SemEval 2016, Task 6-A

### 5.2.2 Approches classiques de détection de posture

Pour accomplir cette tâche, deux approches sont communément suivies. La première, basée sur un dictionnaire de sentiment, suit la supposition qu'un tweet de sentiment positif, devrait être favorable aux entités mentionnées dans le tweet : ainsi, un message envers la cible « Hillary Clinton », de polarité positive, est susceptible d'exprimer un soutien envers la candidate.

Le seconde approche, d'apprentissage supervisé, consiste à apprendre, par un modèle de classification et grâce à un corpus annoté dédié, à labelliser les textes. Les mots fréquemment utilisés dans le corpus d'apprentissage devraient être présents de manière similaire dans le corpus de test.

Dans ce chapitre 5, notre objectif est de montrer que l'utilisation des contextosets permet d'améliorer les performances de classifieurs présents dans la littérature pour la détection de posture.

#### 5.2.2.1 Approche basée sur des ressources linguistiques : le sentiment

Nous proposons une première méthode, nommée SENT-BASE, qui recourt à la ressource bien connue SentiWordNet 3.0 [Baccianella et al., 2010], introduite en chapitre 2. Nous faisons la supposition que les tweets de sentiment valué positivement (resp, négativement) auront la posture *favorable* (resp, *opposée*).

Dans SentiWordNet, chaque terme $n$ peut être présent dans différents *synsets*, des ensembles de synonymes. Chaque élément d'un synset est interchangeable, dans un contexte donné. Soit $\mathrm{S}(n)$ le synset de $i$ mots $s_i$, et contenant le mot $n$. Chaque synset a une valence positive et une négative, notées $s_i^+, s_i^-$.

Soit $\mathrm{S}_t$ l'ensemble des N synsets pris en compte pour un tweet $t$. Nous définissons la valence $v(t)$ :

$$v(t) = \frac{1}{\mathrm{N}} \sum_{s_i \in \mathrm{S}_t} s_i^+ + s_i^- \tag{5.1}$$

Les neutres sont déterminés par un seuil sur la valeur absolue de la valence. Nous avons empiriquement fixé ce seuil à $0,05$ grâce à un corpus d'entraînement.

$$|v(t)| < s_{neutre} \implies posture(t) = \text{NEUTRE} \tag{5.2}$$

Pour des valeurs absolues supérieures au seuil, si $v(t)$ est positive (resp, négative), nous supposons que le tweet est favorable à sa cible (resp, opposé), donc ayant une posture POUR (resp, CONTRE ; FAVOR et AGAINST dans le cadre de SemEval2016).

Les tweets présents dans le corpus disposent déjà d'un label indiquant la cible envers laquelle s'exprime la posture. La détection de cible, ou de sujet de l'opinion, n'est pas un problème trivial mais ne fait pas l'objet de nos travaux de recherche.

#### 5.2.2.2 Approche statistique

Une seconde méthode « classique », à base d'apprentissage automatique, repose sur un SVM (Support Vector Machine) dont les observations sont constituées d'unigrammes : des mots, en singletons, par opposition aux bigrammes et n-grammes. Ces unigrammes sont représentés par une matrice creuse de booléens représentant la présence de certains mots. Nous nommons cette approche SVM-UNIG.

Afin de choisir les meilleurs paramètres, nous recourrons à la méthode de validation croisée. Finalement, suivant l'approche de [Hasan and Ng, 2013], notre choix s'est porté sur un SVM à noyau RBF. Les optimums pour les valeurs de paramètres peuvent varier selon le corpus ; $\mathrm{C} = 100.0$, $\gamma = 0.01$ forment un bon point de départ. Ce sont les valeurs que nous retenons pour SVM-UNIG. L'implémentation est faite en Python, recourant au package Scikit-learn [Pedregosa et al., 2011].

### 5.2.3 Présentation de la difficulté des sociolectes

De manière générale, le traitement automatique de la langue fait face à de nombreuses difficultés. Celles-ci sont renforcées lorsque les messages proviennent des réseaux sociaux : en effet, alors que les premiers modules de détection de thématique, d'opinion ou d'extraction d'informations s'attaquaient à des textes correctement écrits et très formatés, tels que des articles de presse, des romans ou des critiques de film, il faut désormais se pencher sur des messages produits par tout un chacun, des textes qui ne sont pas souvent édités ni relus, et qui ne s'adressent pas à une même audience.

Nous attribuons deux causes à ces difficultés. La première cause correspond au domaine de recherche des « user generated contents », qui inclut les libertés prises quant à l'orthographe et la grammaire, parmi lesquelles se trouve un aspect oralisant. Certains tweets sont écrits comme une exclamation, une intervention orale de son auteur. Il n'y a parfois pas d'orthographe correcte pour des onomatopées ou néologismes ; la faute est parfois voulue, rarement relevée.

La seconde cause est en lien avec l'espace de l'énonciation. Twitter en tant qu'espace social permet la communication entre ses utilisateurs. Les tweets peuvent ainsi être vus parfois comme des messages de personne à personne, parfois comme des affiches, destinées à qui veut bien les lire. Alors que la catégorie « message » connaît des variations sensibles selon le nombre de destinataires explicites, ou si les destinataires sont implicites, il en va de même pour l'aspect « affiche », qui va de la revendication personnelle au communiqué de presse. Enfin, il est des messages qui ne s'adressent à personne, en particulier car certains utilisateurs voient Twitter comme un journal *intime*, bien qu'ouvert à tous.

La différence entre textes normaux et textes « sociaux » est telle que certains parlent des écrits sur Twitter comme d'une langue à part entière. Cela donne lieu à des néologismes, tel *sociolecte*, pour « dialecte de réseaux sociaux » [Farzindar and Roche, 2013].

Comme précédemment mentionné, l'un des grands défis dans l'interprétation des tweets repose sur le non-respect des règles établies de grammaire et d'orthographe : l'émergence de liens sociaux crée une communauté qui se reconnaît par la langue utilisée. De nombreux mots nouveaux émergent, porteurs d'indications importantes sur le contenu sémantique du message. Les éléments les plus visibles en sont les émoticônes et les hashtags, qui sont des manières de souligner respectivement l'émotion et le sujet du message. Par exemple, le hashtag « #demexit » représente la faction d'élus du parti démocrate, quittant le parti, lors de la campagne de 2016. On imagine l'impact d'un tel mot dans l'évaluation de la posture d'un tweet. Ces mots sont encore considérés comme nuisant au sérieux et à la construction du discours. Imagine-t-on une #thèse ironique ☺ ?

D'autres mots n'appartiennent pas à un vocabulaire partagé, mais proviennent de l'application de règles partagées, parmi lesquelles l'allongement de certaines lettres, souvent des voyelles. Par exemple, « laaazzzyymoonnddayyy » (paresseux le lundi) est constitué de l'ajout d'un nombre arbitraire de caractères, selon l'humeur de l'auteur. *Helloooo* semble une meilleure manière de communiquer de l'enthousiasme, qu'un sobre *Hello*. Ce genre de terme est susceptible de n'être présent qu'une seule fois ; des méthodes existent pour trouver la racine « commune » permettant de relier ces instances de mots à leur orthographe originale.

Tous ces éléments dégradent la performance des classifieurs usuels, entraînés sur du texte correct. Nous y voyons un besoin pour une approche dédiée à l'analyse de messages de type *tweet*.

## 5.3 Contextonymes et contextosets

Dans cette section, nous présentons les définitions de contextonymes et contextosets, et nous exposons l'algorithme de construction d'une nouvelle ressource linguistique : un dictionnaire de contextosets.

### 5.3.1 Définitions

Le concept de *Contextonymes* a tout d'abord été introduit par [Hyungsuk et al., 2003], qui écrivait que *« contextually related words are meaningful indicators of a target word's semantic value in a given context »* : la relation contextuelle est un indicateur significatif de la valeur sémantique d'un mot cible dans un co-texte donné. Dans cette étude, les contextonymes sont définis comme « mots contextuellement pertinents pour un mot cible ». Une clarification est nécessaire à propos du terme *contexte* : il s'agit plus ici de co-texte, c'est-à-dire les mots environnants.

Bien que « de niche », les contextonymes ont déjà une histoire et des applications variées : [Ploux and Hyungsuk, 2003] et [Wang et al., 2016] proposent chacun une méthode de traduction statistique, où le mot est remplacé par une « unité minimale sémantique », représentée par une clique de mots obtenue par la méthode d'extraction des contextonymes introduite par [Hyungsuk et al., 2003].

Sur une autre piste, [Şerban, 2013] a extrait des contextonymes à partir de sous-titres de films, représentant des dialogues, afin de proposer une correction de SentiWordNet : les valences de sentiment devraient vraisemblablement pouvoir être propagées le long de l'appartenance à un même contexte d'énonciation.

**Definition 5.3.1.** *Contextonymes.* Deux mots sont contextonymes s'ils surviennent fréquemment ensemble, partageant un même co-texte.

**Definition 5.3.2.** *Contextoset.* Ensemble de mots survenant fréquemment ensemble deux à deux dans un même co-texte.

Les définitions 5.3.1 et 5.3.2 utilisent la notion de co-texte, les mots environnant les termes d'intérêt. En conséquence il n'est pas trivial de déterminer qu'un lien de co-occurrence signifie une contextonymie, ou plusieurs. Pour résoudre ce problème, nous construisons notre approche dans le sens contraire, suivant l'intuition de [Hyungsuk et al., 2003] qui cherche des unités de sens minimal, sous la forme de *cliques* (sous-graphes complets) à partir des liens de co-occurrence. Nous estimons que cette contrainte est trop forte, et que plusieurs cliques très proches, c'est-à-dire partageant un grand nombre de termes, évoquent un même contexte. En conséquence, nous utilisons un algorithme de détection de k-cliques pour proposer les contextosets.

### 5.3.2 Construction d'une ressource linguistique

La méthode de construction nécessite un grand volume de textes, non annotés, représentant le langage visé. La figure 5.2 expose une vue d'ensemble du processus : après un pré-traitement pour extraire les mots des chaînes de caractères, un graphe des co-occurrences des mots est construit, puis nettoyé par filtrages. L'extraction elle-même peut utiliser différentes pistes. Une implémentation précédente cherchait des cliques dans le graphe [Hyungsuk et al., 2003] ; comme discuté précédemment, nous chercherons des k-cliques, pour réduire le nombre de contextosets similaires.


FIGURE 5.2 – Vue générale de la construction des contextosets

#### 5.3.2.1 Ressources requises

Les contextosets sont extraits à partir d'un grand corpus de documents non annotés, qui n'est pas forcément lié au corpus d'entraînement à la détection de posture. Le contenu de ces documents devrait correspondre à des sujets d'intérêt : non pas qu'un corpus générique ne donnerait pas de

contextonymes, mais plutôt qu'il faudrait un corpus suffisamment grand pour favoriser l'apparition de contextes communs, et couvrant toutes les thématiques.

#### 5.3.2.2 Prétraitement des textes

Comme dans toute chaîne de traitement de texte, un nettoyage préalable est nécessaire. Dans cette étape, les caractères sont passés en minuscules, et les mentions d'utilisateurs, symboles spéciaux, et « stopwords » (mots fréquents portant peu de valeur sémantique, tels que « le, la, les, à, de ... ») sont retirés. Quelques abréviations courantes sont transformées en leur version complète (par exemple, « I'm » devient « I am »).

Nous avons considéré l'utilisation d'un lemmatiseur, notamment avec TweetNLP [Owoputi et al., 2013], qui se revendique spécialisé dans le traitement des tweets. Cependant nos tests ne se sont pas révélés convaincants ; par exemple, l'organisation terroriste « ISIS » se trouve lemmatisée en le verbe « is », parmi de nombreuses autres transformations décevantes. Finalement, nous avons préféré ne pas ajouter d'erreurs de lemmatisation, et n'y recourrons pas dans notre méthode.

#### 5.3.2.3 Construction d'un graphe des co-occurrences de mots

**Definition 5.3.3.** *Tweet.* Un tweet $t$ est un ensemble de termes $\{n_i, n_j, ...\}$ obtenus par prétraitement du texte original.

**Definition 5.3.4.** *Co-occurrence.* Les mots $n_1, n_2$ montrent une co-occurrence s'ils sont présents dans un même tweet $t$, et séparés au maximum par $WindowSize - 1$ mots. Autrement dit, il y a co-occurrence entre un mot et les $WindowSize$ mots placés auparavant ; ainsi qu'entre le mot et les $WindowSize$ placés après dans le tweet. $WindowSize$ est un paramètre qui peut prendre la valeur de n'importe quel nombre entier positif, pair. Comme [Şerban et al., 2012], nous avons choisi $WindowSize = 2$ : les mots sont proches dans la phrase, mais incluent plus que la simple juxtaposition.

En recourant à un corpus de tweets prétraités, nous avons construit un graphe des co-occurrences de mots $G = (V, E)$. L'ensemble des nœuds $\{V\}$ est composé du vocabulaire provenant du corpus. L'ensemble des arcs $\{E\}$ représente les liens valués, non dirigés, entre chaque paire de mots apparaissant ensemble (voir définition 5.3.4). Le poids $w_e$ de chaque arc $e$ est le nombre de co-occurrences des mots qu'il relie.

#### 5.3.2.4 Filtrage des mots

Afin de limiter le nombre de mots, nous avons l'intuition qu'il faut filtrer les mots peu présents. Il nous faut aussi retirer les relations trop peu importantes entre les mots : une co-occurrence unique crée un arc qui impacte la détection de contextosets. Cette intuition est renforcée par de précédents travaux sur l'extraction de contextonymie, qui utilisent des filtres sur les fréquences de mots et de co-occurrence [Şerban et al., 2012] : des fréquences trop basses mènent au retrait du nœud ou de l'arc. Ces filtrages simples présentent toutefois des inconvénients :

— Sur les fréquences des mots : le filtrage classique ne prend pas en compte la position des mots dans la phrase ; or l'utilisation d'une fenêtre de co-occurrence met en retrait les mots présents au début ou en fin de tweet, créant un déséquilibre.

— Sur les fréquences des arcs : le filtrage classique ne prend pas en compte le poids de l'association pour chacun des mots, ce qui tend à isoler les mots porteurs de plusieurs sens, mais relativement peu utilisés.

Aussi nous introduisons ci-dessous des définitions pour mettre en œuvre deux mécanismes pour un filtrage plus fin.

**Definition 5.3.5.** *Degré.* Le degré d'un mot $n$ dans un graphe des co-occurrences G est le nombre de mots voisins, i.e. directement connectés à $n$. Nous notons ce degré $d(n)_{\mathrm{G}}$.

Nous pensons le filtrage des mots comme une distinction entre mots légitimes et mots sans sens. Un mot est légitime s'il a été utilisé dans différents contextes, c'est-à-dire utilisé dans des messages différents. Ainsi, le degré du mot dans le graphe aide à répondre à cette interrogation.

De plus, puisque nous nous basons sur une taille de fenêtre $\mathrm{W}indow\mathrm{S}ize$ pour compter les co-occurrences, le degré des mots est impacté par la position du mot dans le texte. Par exemple, « *dogs like to swim in the summers* » et « *dogs usually run very fast* » donnerait aux mots « *dogs* », « *swim* », « *to* », « *in* », « *run* » des degrés de 4, même si « *dogs* » est le seul mot à apparaître dans les deux documents. C'est pourquoi nous introduisons une normalisation du degré du mot par son degré moyen lié à sa position dans le tweet, et obtenons un ratio $\alpha$, représentant la variété de contextes dans lesquels le mot apparaît.

Soit $g_t = (\mathrm{V}_t, \mathrm{E}_t)$ le graphe des co-occurrences des mots d'un tweet $t$. Pour un mot donné $n$, notons 1,...,K l'ensemble des tweets tokenisés contenant $n$. Alors, le degré moyen $\phi$ d'un mot $n$ dû à sa position, est donné par :

$$\phi(n) = \frac{1}{\mathrm{K}} \sum_{j=1}^{\mathrm{K}} d(n)_{g_j} \tag{5.3}$$

Nous définissons alors $\alpha(n)$, le ratio de degré de $n$ dans G par le degré moyen de position pour le mot $n$, en équation 5.4. Un score élevé implique que le mot $n$ apparaît dans une grande variété de contextes. En conséquence, les mots d'un message retweeté à l'identique, mais dont l'orthographe particulière ne serait pas fréquemment reproduite, par exemple : *« dizz movi ezz hoorrble »*, donnerait aux mots le composant des scores de 1, même si ce tweet était réémis 50 fois. Enfin, un mot $n$ sera retiré du graphe G si $\alpha(n) < \alpha_{threshold}$.

$$\alpha(n) = \frac{d(n)_{\mathrm{G}}}{\phi(n)} \tag{5.4}$$

La seconde partie du traitement de filtrage concerne les arcs. Un filtrage classique, éliminant uniquement les arcs de poids faible, éliminerait des contextes potentiellement intéressants mais peu représentés dans le corpus, et favoriserait des co-occurrences concernant les contextes les plus fréquents. Afin de résoudre ce problème, nous introduisons la métrique $\beta$ en équation 5.5, qui consiste en deux ratios de poids d'arcs et de comptage de mots.

$$\beta(e) = \frac{w_e}{c_{n_{1,e}}} + \frac{w_e}{c_{n_{2,e}}} \tag{5.5}$$

Où $w_e$ est le poids de l'arc $e = (n_1, n_2)$, $c_{n_{1,e}}$ et $c_{n_{2,e}}$ représentent le nombre total d'occurrences pour les deux mots $n_1$ et $n_2$ connectés par $e$. Puisque $\beta_e \in ]0,2]$, une valeur approchant 2 implique que cette association est très importante pour les deux mots, tandis qu'une valeur approchant 0 implique une association relativement peu importante.

En retirant les arcs dont les valeurs de $\beta$ sont petites, i.e. si $\beta_e < \beta_{threshold}$, nous éliminons ainsi les relations peu pertinentes, conservant ce qui fera la valeur des contextosets.

#### 5.3.2.5 Extraction des contextosets

Nous extrayons les contextosets via la méthode de découverte des k-cliques proposée par [Palla et al., 2005]. Il s'agit d'une manière de calculer les k-cliques d'un graphe, c'est-à-dire des ensembles de nœuds dans un graphe, dont chaque élément est lié à au moins $k$ autres éléments de l'ensemble. Cela correspond à notre problème car l'approche par cliques maximales (tous les éléments sont liés les uns aux autres) génère de nombreuses cliques ne se différenciant que par un unique mot. De leur côté, les k-cliques ont tendance à agréger de telles cliques maximales. Le choix de $k$ est un compromis pour obtenir suffisamment de contextosets, éviter les couples de mots (une paire de nœuds est une clique), et afin que les contextosets soient suffisamment sensés, focalisés.

## 5.4 Description des contextosets issus de tweets

Cette section expose le résultat obtenu après l'application de l'algorithme de construction de la ressource « contextosets » sur un corpus de tweets. Nous y présentons les détails de l'implémentation, le corpus de tweets non annotés, le graphe des co-occurrences obtenu, ainsi que des contextosets détectés que nous comparons avec d'autres ressources linguistiques. Finalement nous proposons une méthode pour identifier le contextoset correspondant le mieux à un tweet.

### 5.4.1 Détails de l'implémentation

Le prétraitement des textes et la construction du graphe des co-occurrences ont fait l'objet d'une implémentation *ad hoc*, en Python, tirant parti de quelques fonctionnalités proposées par le package *Natural Language ToolKit* [2].

L'implémentation retenue pour détecter les k-cliques est celle de [Palla et al., 2005], disponible dans le package Python NetworkX [3] [Hagberg et al., 2008]. Cette bibliothèque de fonctions de traitement de graphes est à la fois efficace et couvre de nombreuses fonctionnalités ; le maniement des objets graphes est plus aisé et permissif (nommage des nœuds et arcs notamment) que sa rivale, igraph [4].

### 5.4.2 Le corpus *GenTweets*

Nous avons collecté un corpus de tweets, écrits en anglais, via l'API Stream fournie par Twitter. Comme son nom l'indique, cette interface fournit, sur demande, un flux de messages, à l'instant de leur publication. Les deux primitives les plus utilisées sont *sample*, qui envoie 1% des tweets écrits, quelle que soit la langue et le contenu ; et *filter*, qui permet de requérir des messages par auteur, par mot, ou par coordonnées GPS.

Ce corpus, nommé *GenTweets* (« tweets génériques »), consiste en 7 773 089 tweets émis entre le 20 novembre et le $1^{er}$ décembre 2015, en anglais, contenant au moins un mot-clé parmi une liste recouvrant diverses thématiques : la candidate Hillary Clinton, le débat sur l'avortement, la religion, et l'environnement. La liste des mots-clés est disponible en Annexe B.1.

### 5.4.3 Aperçu de la co-occurrence

La figure 5.3 illustre un morceau du graphe des co-occurrences construit à partir du corpus *GenTweets*, autour d'un mot : « support ». La spatialisation choisie, ForceAtlas2, attire les nœuds le long des arcs les liant, et les repousse comme deux aimants de même polarité.

Par construction, « support » apparaît au centre. En bas à gauche, quelques mots semblent groupés : tennessee, trump2016, font référence à la campagne électorale de Donald Trump. Sans surprise, *climate* apparaît non loin de *change*, à droite. Dans la partie supérieure, le groupe formé par les mots *bae*, *naten* et *kanta* illustre notre propos quant aux sociolectes : ces trois mots proviennent de tweets philippins, soutenant des chanteurs de variété par des messages mélangeant anglais et tagalog (la langue majoritaire aux Philippines). De ce graphe, dont seule une partie est visible en figure 5.3, sont extraites des k-cliques, correspondant aux contextosets.

### 5.4.4 Comparaison avec Word2Vec et Wordnet

Afin d'évaluer les relations sémantiques proposées par les *contextosets*, nous les avons comparés avec les résultats de deux autres approches : Word2Vec [Mikolov et al., 2013] et WordNet [Miller, 1995]. Nous avons utilisé le même corpus de 70 millions de mots, *GenTweets*, composé intégralement de tweets, pour les contextosets et pour Word2Vec.

2. www.nltk.org/
3. Documenté sur : `networkx.github.io`
4. Accessible sur `igraph.org/python/`

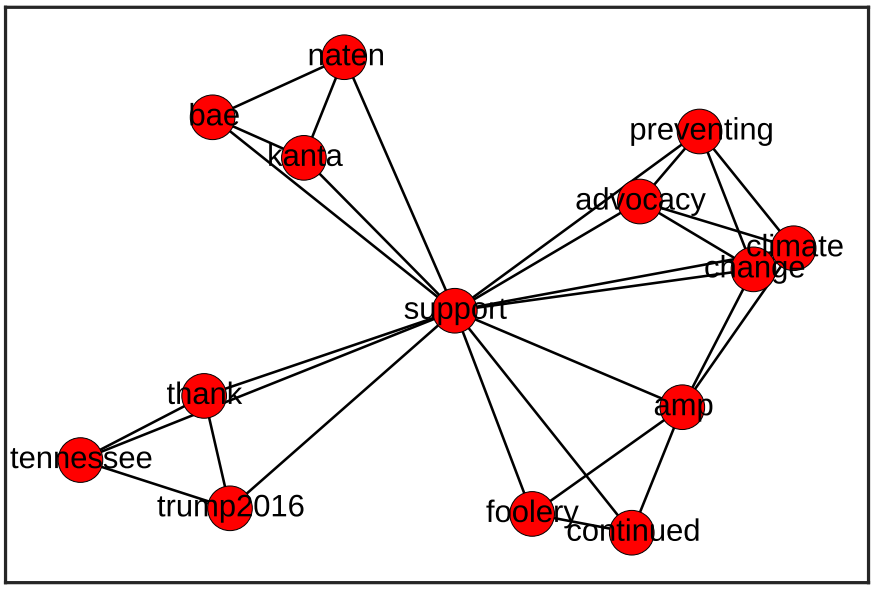



FIGURE 5.3 – Co-occurrence de mots autour de *support*

Nous avons obtenu les *word embeddings* en entrainant un modèle Word2Vec, suivant l'approche en sac de mots et des vecteurs de dimension 100. La table 5.1 présente les mots proches du mot « *support* » : nous y montrons quelques synsets de WordNet, les mots voisins selon Word2Vec, et les cliques de contextonymes.

Parmi les résultats proposés par Word2Vec, seuls quelques mots semblent reliés à notre sujet. Les catégories grammaticales ne sont pas considérées, et nous n'avons aucune idée des relations entre les mots proches de la cible : peut-être sont-ils assez éloignés les uns des autres. Par exemple, une seconde requête montrera que « *respect* » n'est pas inclus dans les mots proches de « *organize* », alors qu'ils apparaissent tous deux près de « *support* ».

Les ensembles de synonymes, ou synsets, de WordNet, sont parfois très nombreux. Par contre ils n'incluent pas le vocabulaire spécifique à Twitter : même le très populaire « *LOL* » est exclu de ce dictionnaire. Cela ne décroît pas la qualité de WordNet, mais il faut reconnaître qu'il n'est pas adapté aux médias sociaux.

Les contextonymes, à l'instar de Word2Vec, ne font pas la distinction entre les catégories grammaticales et peuvent inclure n'importe quel terme présent dans le corpus. De plus, ils convoient assez efficacement la signification d'un mot, et correspondent évidemment aux sujets originaux du corpus.

#### 5.4.4.1 Exploitation des contextosets

Afin de tirer profit de la ressource linguistique constituée, nous proposons un mécanisme d'obtention du ou des contextosets correspondant le mieux à un tweet. Tout d'abord, nous obtenons une liste de contextosets candidats, en utilisant comme critère le nombre de mots partagés entre chaque contextoset et le tweet. Il est possible d'avoir plusieurs contextosets avec le même nombre de mots présents dans le tweet.

Soit C l'ensemble des contextosets $c$ générés à partir de *GenTweets*. Alors, pour un tweet $t$, composé des termes $n_t$, et pour chaque contextoset $c$ composé des termes $n_c$, l'ensemble des meilleurs contextosets $B_t$ est donné par l'équation 5.6.

TABLEAU 5.1 – Word Embeddings, contextosets et WordNet synsets pour les mots proches de *support*

| **Method : Word Embeddings** |
|---|
| supporting, supported, supports, respect, vote, encourage, voting, voted, organize, helping |
| **Method : Contextosets** |
| (support, continued, foolery),<br>(climate, support, advocacy, preventing, change),<br>(support, bae, naten, kanta),<br>(support, tennessee, thank, trump2016) |
| **Method : Synsets** |
| (documentation, support)<br>(support, keep, livelihood, living, bread and butter, sustenance)<br>(support, supporting)<br>(accompaniment, musical accompaniment, backup, support)<br>(support, financial support, funding, backing, financial backing)<br>(support, back up)<br>(back, endorse, indorse, plump for, plunk for, support)<br>(hold, support, sustain, hold up)<br>(confirm, corroborate, sustain,substantiate, support, affirm)<br>(subscribe, support)<br>(corroborate, underpin, bear out, support)<br>(defend, support, fend for)<br>(patronize, patronise, patronage, support, keep going)<br>(digest, endure, stick out, stomach, bear, stand, tolerate, support, brook, abide, suffer, put up) |

$$\mathrm{B}_t = \{c \mid max(|\{n\}_c \cap \{n\}_t|), \forall\ c \subset \mathrm{C}\} \tag{5.6}$$

Par la suite, les modules de classification usuels (« -BASE ») sont alimentés par le tweet $t$ tandis que leurs versions améliorées (« -CTXT ») reçoivent à la fois $t$ et $\mathrm{B}_t$.

Dans cette section nous avons introduit une nouvelle ressource linguistique, les contextosets. Nous avons donné les détails de l'implémentation ainsi que les spécifications d'un corpus de tweets, non annotés, servant de donnée d'entrée. Nous avons ensuite montré quelques contextosets, et nous les avons comparé à d'autres ressources bien connues : les synsets proposés par WordNet, et la similarité entre termes introduite par Word2Vec. Nous allons désormais utiliser ces contextosets dans notre tâche de détection de posture, en introduisant une façon originale d'exploiter cette ressource construite sur un corpus non supervisé.

## 5.5 Évaluation des contextosets pour la détection de posture

Nous avons choisi la tâche proposée par la campagne SemEval2016 pour plusieurs raisons. D'une part, il s'agit d'une composante intéressante de la détection d'opinion, une fonctionnalité désirée par de nombreux analystes de réseaux sociaux. D'autre part, la mise à disposition tant de corpus d'entraînement et de test que de concurrents permet de poser le problème et d'évaluer objectivement la performance et la qualité des méthodes déployées.

### 5.5.1 Le corpus SemEval

Nous évaluons nos classifieurs de posture sur le corpus SemEval2016-task6 [5]. Nous nous focalisons surtout sur la sous-tâche A [6], qui inclut une division entraînement/test, sur chacun des cinq sujets suivants :

— *athéisme*, où la posture « pour » correspond à la promotion de l'athéisme ; « contre » signale un contenu favorable à la religion ;

— *le changement climatique est un vrai problème*, où la posture « pour » labellise les messages écologistes, « contre » marque les climato-sceptiques ;

— *féminisme*, où la posture « pour » marque les messages féministes ou égalitaires ; les textes sexistes ou opposés aux féministes sont signalés par un « contre » ;

— *Hillary Clinton* où ses soutiens sont marqués « pour », les opposants « contre » ;

— *légalisation de l'avortement*, dont les partisans de son autorisation sont signalés par un « pour » ; le mouvement pro-life, anti-avortement par un « contre » ;

Pour chacune des cibles précédentes, nous disposons aussi de tweets « neutres », correspondant soit à l'absence d'éléments permettant la détermination de la posture, soit à l'absence d'éléments faisant référence à la thématique liée à la cible. Le premier point s'explique notamment par la présence de tweets informatifs ou flous. Le second point est lié à la méthode de collecte des tweets, par hashtags ensuite retirés du corpus : parfois seul le hashtag final permettait de donner un sens humainement compréhensible au message.

Le corpus sur les cinq thématiques comprend 2 914 tweets dans la partie entraînement ; la partie test est composée de 1 250 tweets équitablement répartis sur les 5 thèmes ; elle n'était pas disponible lors de la compétition, conclue en janvier 2016. Les participants étaient évalués via une métrique officielle, le *OfficialScore* (voir équation 5.11). Les deux jeux de données, d'entraînement et de test, sont désormais librement téléchargeables.

### 5.5.2 Exploitation des contextosets pour améliorer la détection de posture

Nous détaillons ici l'implémentation des versions améliorées, utilisant les contextosets, pour chacune des deux approches proposées précédemment : en premier, celle qui utilise SentiWordNet comme dictionnaire de sentiment ; puis l'approche statistique, qui apprend à partir du corpus d'entraînement.

#### 5.5.2.1 Approche par dictionnaire de sentiment

La prédiction de sentiment peut être améliorée si elle considère aussi la relation de contextonymie lors de la sélection des synsets. La méthode à base de sentiment, exploitant les contextonymes, est nommée SENT-CTXT.

Son entrée est constitué du tweet $t$ tokenisé ainsi que du (ou des) contextoset correspondant, $B_t$. La méthode choisit pour chaque mot le meilleur des synsets de SentiWordNet (associant un ensemble de mots synonymes, et une valence), selon le nombre de mots partagés entre le mot-token présent dans le tweet, les membres du contextoset, et le synset.

Si deux synsets sont en compétition, leurs champs de description (le champ *glossary* dans le dictionnaire WordNet constitué de quelques mots ou phrases permettant au lecteur d'affiner sa compréhension du sens du synset) alimente le choix, et permet de finalement retenir l'unique meilleur synset de SentiWordNet : celui qui partage le plus de termes avec le tweet.

Finalement, la valence du tweet est calculée comme pour SENT-BASE, et permet d'afficher un label de posture. Les contextosets permettent ici de lever les ambiguïtés en choisissant plus intelligemment les synsets, donc les valences, à prendre en compte.

5. SemEval2016-task6 est librement téléchargeable à l'adresse suivante : http ://alt.qcri.org/semeval2016/task6/

6. La sous-tâche B, focalisée sur Donald Trump, ne comprenait pas de tweets annotés permettant l'entraînement, afin d'encourager les méthodes non-supervisées.

#### 5.5.2.2 Approche statistique

Pour améliorer la brique de classification supervisée introduite précédemment (SVM-UNIG), nous suivons l'intuition suivante pour concevoir SVM-CTXTS : peut-être que les contextosets eux-mêmes sont de bons indicateurs de la posture des tweets. Plutôt que de se baser sur la présence des unigrammes, il est possible de représenter un texte par les contextosets qu'il évoque, et donc par des booléens indiquant la présence de ces contextosets.

Nous pensons que cette méthode est assez sensible à la taille du corpus. Alors que nous avons obtenu 6 278 contextosets à partir du corpus *GenTweets*, il est vraisemblable de ne constater l'occurrence d'un contextoset qu'une fois de temps en temps, pas assez pour apprendre à partir d'un corpus de petite taille (si l'on veut 10 textes-exemples par contextoset, il faudrait 62 000 documents pour l'apprentissage de la posture).

Une seconde piste pour exploiter les contextosets dans le cadre d'un apprentissage supervisé, que nous nommons SVM-EXP, consiste à répondre au problème de la taille, trop courte, des tweets, en les enrichissant avec les termes composant les contextosets évoqués dans le tweet. Les messages sont étendus en y ajoutant les éléments (les mots) de $\mathrm{B}_t$, comme défini en équation 5.7 : il s'agit de l'union des termes du messages avec les éléments du ou des meilleurs contextosets.

$$\mathrm{E}_t = \{n\}_{\mathrm{B}_t} \cup \{n\}_t \tag{5.7}$$

Après quoi le tweet tokenisé, représenté par une liste de termes, est analysé par un classifieur similaire à SVM-UNIG (mais entraîné sur la version étendue des tweets) pour déterminer la posture. Chaque message est ainsi représenté par un ensemble plus grand d'unigrammes, ce qui aide à déterminer son contexte et donc, sa posture.

### 5.5.3 Mesures d'évaluation

Chaque analyseur de posture peut être vu comme constitué de plusieurs classifieurs mono-label. Étant donné un texte, le classifieur d'une posture $s$ peut s'activer (être *positif*), et lui attribuer un label qui correspond à sa vérité, ou non. Les classifieurs correspondants aux autres postures ne s'activent alors pas (sont *négatifs*). Ce texte est alors compté comme *vrai* s'il correspond à la vérité terrain, ou *faux* en cas d'erreur.

Pour calculer des métriques d'évaluation des résultats, un ensemble d'observations dont les labels sont connus (on parle alors de *gold labels*, ou de vérité terrain), appelé *test set*, est nécessaire.

En conséquence, chaque échantillon peut générer quatre résultats possibles envers chaque posture $s$ : *TP* représente le nombre de vrais positifs (*true positives*), signifiant que le classifieur a correctement déterminé une quantité TP d'échantillons comme porteurs de la posture $s$. *TN* représente les vrais négatifs, soit le nombre de fois où le classifieur a correctement reconnu que l'échantillon ne portait pas la posture $s$. Enfin, l'erreur est représentée par *FN*, les faux négatifs : le classifieur ne s'est pas activé alors qu'il aurait dû ; et *FP* (faux positifs) : le classifieur a affirmé la présence de la posture $s$, alors que ce n'est pas le cas.

La première métrique, la *précision*, est définie dans l'équation 5.8 et représente la fraction de *vrais positifs*, i.e les observations correctement classifiées comme étant de posture $s$. Lorsque le classifieur affirme un résultat, est-il vrai ? Une autre métrique, le *rappel*, définie en équation 5.9, représente la fraction d'observations dont le vrai label est $s$, classifiées comme étant $s$. Est-ce que le classifieur trouve tous les échantillons marqués $s$ ?

Ces deux mesures peuvent être combinées de plusieurs manières : leur moyenne simple s'appelle *fiabilité*, mais ne représente pas bien la qualité du classifieur lorsque la répartition des classes est déséquilibrée. La moyenne harmonique, ou mesure $\mathrm{F}_1$, définie en équation 5.10, est communément utilisée pour évaluer la qualité de la prédiction des classifieurs.

$$P_s = \frac{TP_s}{TP_s + FP_s} \tag{5.8}$$

$$R_s = \frac{TP_s}{TP_s + FN_s} \tag{5.9}$$

$$F_1(s) = 2\frac{P_s R_s}{P_s + R_s} \tag{5.10}$$

Pour comparer les résultats des différents compétiteurs à SemEval, une métrique officielle a été proposée. Elle consiste en la moyenne des $F_1$ mesures pour les postures positives (F pour Favor) et négatives (A pour Against), n'incluant donc pas la posture *neutre*, correspondant dans la taxonomie de SemEval, à la fois aux discours objectifs, i.e n'exprimant pas de soutien, et aux discours hors sujets. Cet *OfficialScore* n'est pas directement comparable aux précision, rappel et $F_1$ mesures exprimés précédemment ; le meilleur score est le plus élevé.

$$\textit{OfficialScore} = \frac{1}{2}\ (F_1(F) + F_1(A)) \tag{5.11}$$

### 5.5.4 Résultats

Le tableau 5.2 contient les résultats de nos expériences. *Précision* et *Rappel* sont des valeurs moyennes sur les trois postures, pour chacune des cibles avec un modèle dédié. $F_1$ est la moyenne des $F_1$-mesures.

TABLEAU 5.2 – Comparaison entre les algorithmes proposés pour SemEval TaskA

| | Algorithme | Précision | Rappel | $F_1$ | OfficialScore |
|---|---|---|---|---|---|
| Sent | SENT-BASE | 0.41 | 0.30 | 0.31 (-) | 0.32 (-) |
| | SENT-CTXT | 0.43 | 0.35 | 0.37 (+0.06) | 0.37 (+0.05) |
| Stat | SVM-UNIG | 0.63 | 0.62 | 0.62 (-) | 0.62 (-) |
| | SVM-CTXT | 0.58 | 0.61 | 0.58 (-0.04) | 0.58 (-0.04) |
| | SVM-EXP | 0.73 | 0.65 | **0.67** (+0.05) | 0.65 (+0.03) |

SENT-BASE se révèle insatisfaisante, avec une $F_1 = 0.31$. Cependant, sa version améliorée SENT-CTXT obtient un meilleur résultat, à $F_1 = 0.37$. Ces faibles scores sont dus à l'hypothèse initiale (lien entre sentiment et posture) qui n'est sans doute pas vérifiée.

De plus, le corpus (entraînement et test) inclut des messages mentionnant d'autres entités que l'entité-cible : par exemple, l'ensemble de messages dont la posture doit être calculée vis-à-vis de Hillary Clinton inclut des messages de polarité positive envers d'autres candidats opposés à Hillary. Soutenir un opposant est assimilé à une opposition à la cible, résultant en une posture négative. L'approche basée sur le sentiment réagit mal à ce genre de piège. Nous pensons que pour traiter ce problème, le système devrait aussi inclure une brique de détection du sujet de l'opinion. En d'autres termes, la cible du sentiment d'un tweet peut différer de la tâche de détection de posture : l'approche par sentiment, bien qu'intuitive, ne convient pas ici.

SVM-UNIG est une meilleure *baseline*, parce qu'il profite du jeu de données d'entraînement (à l'opposé de SENT-BASE), et obtient un score honorable $F_1 = 0.62$. On pourrait objecter que la meilleure performance sur cette tâche (*OfficialScore* de 0.68) a été obtenue par un classifieur similaire, un SVM dont les caractéristiques recouvrent aussi les n-grammes de caractères, exposé comme une baseline par les organisateurs du concours SemEval [Mohammad et al., 2016].

La faible performance de SVM-CTXT est, selon nous, liée à la faible taille du corpus d'entraînement. L'ensemble de messages a en effet peu d'éléments pour couvrir le vocabulaire nécessaire pour matcher les possibilités des contextosets, poussant SVM-CTXT à se prononcer pour des contextosets qu'il n'a jamais vus auparavant.

Finalement, SVM-EXP montre une amélioration par rapport à sa version initiale (SVM-UNIG), atteignant $F_1$ = 0.67 (*OfficialScore* = 0.645). Ce n'est certes pas mieux que les meilleurs sur cette tâche, mais c'est comparable. Pour mieux analyser l'amélioration introduite par les contextonymes, la table 5.3 compare SVM-UNIG et SVM-EXP en termes de précision, rappel et $F_1$ pour chacune des trois postures. SVM-EXP est nettement meilleure sur les postures *Positive* et *Négative*, ce qui se répercute sur le score officiel. Par nature, les tweets de posture *Neutre* sont plus dispersés ; il semble que l'ajout au tweet des termes du meilleur contextoset ne permet pas de rapprocher les tweets neutres entre eux.

TABLEAU 5.3 – Comparaison par posture pour SVM-UNIG et -EXP

| Posture | Précision | | Rappel | | $F_1$ | |
|---|---|---|---|---|---|---|
| SVM- | UNIG | EXP | UNIG | EXP | UNIG | EXP |
| Négative | 0.74 | 0.87 | 0.71 | 0.68 | 0.73 | **0.76** |
| Positive | 0.54 | 0.44 | 0.47 | 0.66 | 0.51 | **0.53** |
| Neutre | 0.40 | 0.24 | 0.53 | 0.39 | **0.45** | 0.30 |
| Total | 0.63 | 0.73 | 0.62 | 0.65 | 0.62 | **0.67** |

Le tableau 5.4 compare notre meilleur algorithme aux meilleures propositions soumises durant l'évaluation SemEval. Notre méthode SVM-EXP aurait été classée $6^{eme}$ parmi les 19 compétiteurs : nos résultats sont acceptables par rapport à la concurrence. Cependant, et c'est valable pour tous les participants, il faut reconnaître que les scores ne sont pas très élevés et que la fiabilité d'une prédiction est trop faible pour la rendre exploitable seule.

TABLEAU 5.4 – Comparaison avec les compétiteurs de SemEval, selon l'*official score*

| Algorithme | A#1 | A#2 | A#3 | A#4 | A#5 | SVM-EXP | A#6 | ... |
|---|---|---|---|---|---|---|---|---|
| *OfficialScore* | 0.678 | 0.673 | 0.668 | 0.658 | 0.656 | 0.645 | 0.636 | ... |

L'organisateur, [Mohammad et al., 2016], propose une analyse des résultats. Le meilleur algorithme au classement (A#1, MITRE) utilise deux réseaux de neurones récurrents (RNNs). Le premier réseau choisit les meilleurs hashtags sur un corpus de tweets non annotés, et le second estime la posture. Le second compétiteur, (A#2, pkudlab), utilise à la fois un réseau convolutif profond (CNN) ainsi qu'un ensemble de règles. Cette approche hybride n'est entraînée que sur l'ensemble de tweets d'entraînement. Nous n'avons aucune information sur la technique suivie par (A#3, TakeLab).

En conclusion de ce comparatif, il semble que le type de classifieur choisi importe finalement assez peu ; les différences portent surtout sur l'effort placé pour élargir la base d'apprentissage (souvent, par sélection de hashtags *espérés* discriminants), et sur les pré-traitements linguistiques, qui peuvent parfois permettre, tels nos contextonymes, de prendre en compte des termes nouveaux et de leur associer un sens.

## 5.6 Discussion

La détection de posture dans un tweet est une tâche difficile, à cause du faible nombre de mots composant le tweet, des orthographes inventives, des règles grammaticales évolutives et de l'usage social des mots. La thématique est souvent implicite, et le sujet d'une opinion lui-même n'est pas toujours explicitement mentionné.

Dans le domaine de la désambiguïsation sémantique, la contextonymie et les contextosets aident à répondre à certaines de ces difficultés, en ajoutant de la clarté au sens du tweet, en faisant le lien avec les contextes usuels d'emploi des termes du tweet. En outre, il est *a priori* possible, de générer des contextosets à partir de messages écrits dans n'importe quel langue ou dialecte, si l'on dispose de quelques outils basiques (i.e. la phase de prétraitement, et la tokenisation), ainsi que d'un grand nombre de textes. Pour nos expériences, nous avions plus de cinq millions de messages.

Afin de montrer l'intérêt des contextosets, nous avons proposé de mesurer leurs effets sur une tâche d'évaluation internationale, la détection de posture dans le cadre de la campagne SemEval 2016. Nous avons introduit deux méthodes naïves : un analyseur de sentiment, basé sur SentiWordNet, et un classifieur de texte, basé sur un SVM. Nous pensons que les contextosets ont un grand potentiel : dans les deux approches, l'exploitation des contextosets augmente la qualité du résultat, évalué par la $F_1$-mesure, bien que l'approche par sentiment ne semble pas adaptée à cette tâche.

Même si l'analyseur de sentiment n'a pas pu prédire correctement certains messages, il reste le seul, par son approche générique, à ne pas nécessiter la constitution coûteuse de corpus d'entraînement spécifique à des tâches ou cibles très particulières. Sur des messages reçus sans annotation de cible, il est calculable et son résultat est le plus souvent vrai ; un module dédié à une cible donnée serait meilleur sur sa cible, mais ne produirait que du bruit sur les autres thématiques.

De plus, nous soulignons le besoin d'avoir un module d'extraction de la cible intrinsèque du tweet, par opposition à l'annotation de la posture vis-à-vis d'une cible proposée par la tâche, et donc extrinsèque. Cette approche permet de traiter n'importe quel texte sans *a priori* de cible, et donc de savoir de quoi parlent les gens sur le réseau social (à l'opposé, en fixant la cible, nous étudions *comment* les gens parlent).

L'approche statistique, fournissant des résultats déjà acceptables, profite de la contextonymie pour clarifier les tweets où il y a de l'implicite, du doute. Outre le gain en temps de calcul, cela pourrait permettre d'améliorer la qualité et de gérer un score d'ambiguïté, ou de confiance, en la prédiction. Cette approche nécessitant un corpus d'apprentissage, elle n'est valable que pour des thèmes susceptibles de se maintenir dans le temps.

## 5.7 Synthèse

L'analyse automatique de textes est un domaine difficile ; sur des tweets, la performance d'un détecteur de posture reste basse. Dans ce chapitre, nous avons construit une ressource linguistique, les contextosets, qui incorporent des mots nouveaux, en usage sur certaines thématiques, à partir d'un corpus non annoté. Nous avons repris la notion de contextonymie de la littérature ; nous avons adapté la méthode d'extraction existante en identifiant des manques dans la procédure de filtrage du graphe des co-occurrences, et en corrigeant ces manques. L'adaptation de classifieurs de posture pour exploiter la ressource obtenue nous a permis d'améliorer une méthode de détection de posture, dans le cadre de la campagne SemEval 2016. Ces travaux ont donné lieu à une publication [Gadek et al., 2017a].

Cette tâche a permis de mettre en place une chaîne de traitement du texte et d'identifier des difficultés de traitement, notamment sur l'extraction d'opinion. Dans un second temps, notre chaîne permet d'enrichir l'analyse des profils d'utilisateurs ainsi que la caractérisation des groupes d'influence par des aspects textuels et sémantiques. Il s'agissait ici d'un premier niveau de l'analyse des réseaux sociaux : l'analyse des messages échangés.

# CHAPITRE 6

## Caractérisation des acteurs-clés : scores, profils et rôles

Le chapitre précédent porte sur l'analyse des textes des messages échangés sur un réseau social ; en parcourant ces messages, une question vient naturellement à l'esprit : qui est l'auteur parmi la foule des utilisateurs ? La mesure de l'impact, de l'influence d'un compte, suscite des applications directes notamment en marketing : contre rétribution, un compte influent peut se transformer en panneau publicitaire. L'utilisation des médias sociaux pour la mesure de l'opinion publique, bien que biaisée, gagne en fiabilité si le sondeur peut relier opinions et catégories d'individus. Le fournisseur de service, c'est-à-dire les administrateurs du réseau social, sont intéressés par l'identification de robots, afin d'éviter de fournir un terreau fertile aux spammeurs par exemple. Aussi, dans ce chapitre, nous scindons cette interrogation en trois. La première question porte sur l'impact que peut avoir un utilisateur par ses messages, et son potentiel de diffusion. Bien que partiellement traité dans l'état de l'art, nous proposons ici quelques mesures, permettant de représenter et quantifier « l'influence » d'un compte utilisateur.

La seconde question concerne le comportement des auteurs des messages. Plusieurs taxonomies sont en concurrence sur ce domaine : une distinction entre comptes personnels, automatisés ou institutionnels peut être faite ; une autre distinction sur le milieu socio-économique dont est issu l'utilisateur pourrait être envisageable : chaque métier a des besoins spécifiques, qui peuvent différer selon les cas précis. L'analyse d'un réseau social requiert un ensemble de statistiques décrivant le *comportement* d'un compte, ouvrant ainsi la voie à des fonctionnalités plus avancées : requête d'utilisateurs similaires selon un aspect ou une dimension donnée ; catégorisation large d'utilisateurs et calcul de statistiques à un niveau macro ; analyse temporelle d'un compte en particulier, à la recherche de patrons ou d'anomalies.

La troisième question concerne la qualité du lien social reliant un compte à son environnement. Intuitivement, cette position du nœud-compte utilisateur dans le réseau-graphe social résulte des actions effectuées, donc du comportement ; cependant la réponse renvoyée par les voisins d'un compte dépend de nombreux facteurs. Finalement, la position sociale caractérise bien l'identité d'un compte (vu comme un nœud d'un graphe), et ne peut être déduite uniquement de l'activité du compte.

Pour répondre à ces trois grands axes, ce chapitre est divisé comme suit : la section 6.1 contextualise le problème et introduit des définitions autour de la notion de compte utilisateur ; la section 6.2 introduit un exemple à vocation illustrative, puis la section 6.3 propose une réponse au besoin de mesure de l'impact d'un utilisateur. Ensuite, la section 6.4 détaille la construction des *profils* des individus, afin d'analyser les comportements dans une grande population. La section 6.5 propose une piste pour étiqueter, de façon automatique, le *rôle* adopté par un utilisateur, en lien avec sa position dans le graphe social. Finalement, le chapitre se clôt par une synthèse en section 6.6.

## 6.1 Caractérisation d'un compte influenceur

Cette section décrit les fonctionnalités d'analyse des comptes, précise le découpage conceptuel entre scores d'influence, profils de comportement et rôle social, et discute des données requises pour procéder à la caractérisation recherchée.

### 6.1.1 Fonctionnalités liées à l'analyse des comptes

D'après le chapitre 3, et malgré les travaux portant spécifiquement sur la détection de spam ou sur les modèles de propagation de l'information, il reste un verrou portant sur l'établissement d'une représentation des comptes utilisateurs qui permette de répondre aux besoins provenant de différents métiers :

— identification d'acteurs-clés ;

— détermination de l'impact potentiel d'un compte ;

— requête d'utilisateurs similaires selon un aspect ou une dimension donnée ;

— catégorisation large d'utilisateurs ;

— calcul de statistiques à un niveau macro ;

— analyse temporelle d'un compte en particulier ;

— recherche de patrons ou d'anomalies.

Ces pistes s'alimentent d'une diversité des données disponibles pour mener à bien cette analyse. Valeurs numériques, messages émis, images de profil, champs textes de biographie, liens vers des groupes, des pages Web, relations sociales et interactions : il est possible d'extraire différents modèles du même réseau, qui se complètent et améliorent la vue globale du paysage informationnel.

### 6.1.2 Trois axes de représentation de l'utilisateur

Pour répondre à la variété des besoins recouvrant la simple question « qui ? » en utilisant au mieux la diversité des données disponibles, nous avons décidé de calculer trois représentations d'un compte.

Tout d'abord, un volet couvre la notion de **mesure de l'influence** : une valeur, facile à représenter, permettant de comprendre l'importance qu'a un compte. La revue de ces indicateurs en chapitre 3 laisse de la place pour des **scores** ne fonctionnant pas en boîte noire (tel le Klout [1]), et plus faciles à appréhender que le simple nombre d'abonnés ou la quantité de messages émis.

Dans un second temps, un volet s'attaque au **profilage** du comportement d'un compte, plus complet qu'un simple compteur d'activité. Nous souhaitons calculer ici de nombreuses valeurs couvrant le spectre des actions possibles, représentant l'utilisateur par un vecteur réel, ce qui ouvre la voie à la catégorisation et à la recherche de valeurs similaires. Ces valeurs seront regroupées en une entité, que nous nommons **profil**.

Enfin, un troisième volet s'attaque à la **position** du compte sur le graphe représentant le réseau social. Bien qu'il y ait une relation entre les actions effectuées (volet comportemental) et les éléments servant à la construction du graphe, nous pensons que la position amène des indications supplémentaires. En effet, il ne suffit pas d'avoir un degré élevé (beaucoup d'actions effectuées) pour être un élément central. Une position dans le graphe étant, par nature, complexe à visualiser, nous la résumerons par l'attribution d'une étiquette de position-type, que nous nommons **rôle** par la suite.

Pour clarifier ces différences, la figure 6.1 positionne les profils-types, scores d'influence et rôles. En vert, l'entité « compte utilisateur » accomplit des actions, qui génèrent des données. Ces données (en violet) sont vues de deux façons : comme des messages, ou comme des relations entre comptes. Leur exploitation génère des informations (en bleu foncé) caractérisant les comptes utilisateurs.

1. https ://klout.com

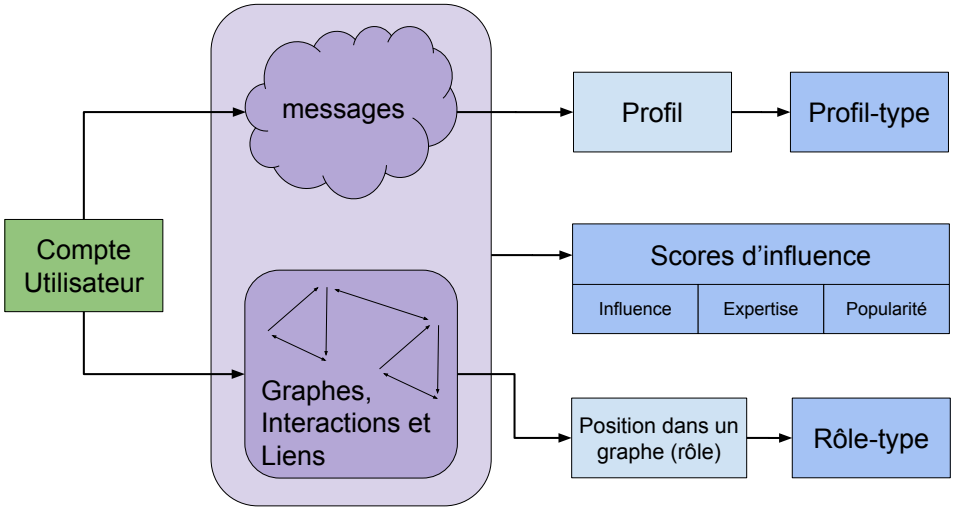



FIGURE 6.1 – Schéma conceptuel du profil, du rôle et des scores d'influence

Le profil, calculé à partir des publications d'un auteur, regroupe la totalité des caractéristiques d'action du compte, y compris sur le style des messages, la temporalité, le contenu. L'information nécessaire provient totalement des actions accomplies par l'auteur ; en cela le profil est intrinsèque à un compte utilisateur. Le profil-type permet de résumer le profil en le rattachant à une catégorie.

Enfin, les scores d'influence (influence, expertise et popularité) peuvent être calculés soit à partir des messages (grâce aux méta-données qu'ils contiennent), soit à partir du graphe (favorisant des positions centrales) : les deux sources d'informations sont nécessaires. L'influence se mesure comme une réalisation d'un potentiel interne : il faut que le profil existe (émette des tweets pertinents, soit une référence à une entité connue) pour mesurer l'adhésion qu'il suscite auprès de son environnement.

La construction du graphe social permet de révéler bien plus que le simple nombre de personnes mentionnées contenus dans le profil. Le graphe révèle la position de chacun. Cette position et ce graphe sont difficiles à appréhender : au-delà de trente nœuds, la lecture d'un graphe devient difficile ; au-delà de trois cent, l'affichage pose problème. C'est alors que le rôle permet d'établir une taxonomie des types de position adoptées par chacun des nœuds du réseau, et donc rendre accessible l'information désirée. À l'opposé du profil, le rôle est calculé « uniquement » sur les données d'interaction sociale, résultant tant des contacts pris par l'auteur auprès de son environnement, que des actions accomplies par l'environnement à son encontre. Le rôle porte ainsi une dimension extrinsèque, sociale.

### 6.1.3 Quantité et qualité de la collecte des données nécessaires

Concernant l'analyse de réseaux sociaux, tel Twitter, il est irréaliste de s'attendre à obtenir une vue complète, intégrale du réseau. D'une part, certaines données, comme le nombre d'abonnés d'un compte, sont faciles d'accès et permettent de rapidement mesurer la taille de l'audience directe d'un utilisateur.

D'autre part, l'analyse du comportement nécessite, par nature, l'observation de comptes en continu sur une période relativement longue. À la contrainte de la durée s'ajoute celle de la quantité : afin d'établir des catégories, des regroupements de comportements similaires, une population est nécessaire. La création d'un corpus comportemental doit alors porter sur plusieurs centaines ou milliers de comptes.

Enfin, l'établissement de positions-types dans un graphe nécessite d'obtenir un graphe : la collecte d'informations doit donc porter sur des comptes globalement liés entre eux. Pour chacun des comptes analysés, l'analyse requiert l'intégralité des actions qu'ils émettent et reçoivent : une attention toute particulière doit porter sur la « frontière » de la population de comptes collectés, c'est-à-dire

les actions concernant la population de comptes visés, mais effectuées par des comptes extérieurs à cette population.

## 6.2 Construction d'un exemple à vocation illustrative : *CinéTweets*

Afin d'illustrer notre propos, nous adossons nos modèles théoriques à un exemple artificiel qui permet de mieux appréhender les techniques introduites. Constitué d'une dizaine de messages, ce jeu de données ne reprend pas toutes les caractéristiques réelles d'un corpus de tweets, mais il en contient l'essence. Nous l'appelons *CinéTweets*, en raison de la thématique choisie.

### 6.2.1 Les messages

TABLEAU 6.1 – Ensemble de tweets pour l'exemple illustratif *CinéTweets*

| Auteur | Message | Interaction | Thème |
|---|---|---|---|
| 1 | hello @2 viens sur mon site : http :url | ME | Autre |
| 1 | hello @3 viens sur mon site : http :url | ME | Autre |
| 1 | hello @4 viens sur mon site : http :url | ME | Autre |
| 6 | un cinéma @5 ? | ME | Cinéma |
| 5 | @6 : quel film ? #Film, #Film1 ou photo_Film_2 ? | RE | Cinéma |
| 5 | allons au cinéma @4 ! @6 vient aussi | ME | Cinéma |
| 4 | @5 : ok | RE | Autre |
| 4 | super #Film, à la prochaine @6 | ME | Cinéma |
| 5 | RT @4 : super #Film, à la prochaine @6 | RT, ME | Cinéma |
| 6 | RT @4 : super #Film, à la prochaine @6 | RT, ME | Cinéma |
| 5 | la bande-annonce de photo_Film_2 :) ! #Film2 dès demain ! | | Cinéma |
| 4 | RT @5 : la bande-annonce de photo_Film_2 :) ! #Film2 dès demain ! | RT | Cinéma |
| 4 | au fait, @8, tu devrais aller voir #Film | ME | Cinéma |
| 5 | le théâtre c'est bien aussi | | Théâtre |
| 4 | RT @5 : le théâtre c'est bien aussi | RT | Théâtre |
| 7 | RT @5 : le théâtre c'est bien aussi | RT | Théâtre |
| 8 | RT @5 : le théâtre c'est bien aussi | RT | Théâtre |
| 6 | @7 : cinéma :-) theââââatre :-( | RE | Cinéma |
| 7 | viens @8, allons au théâtre | ME | Théâtre |

La table 6.1 regroupe des messages artificiels, rassemblés par ligne narrative pour la clarté du propos ; la plupart d'entre eux portent une interaction, qui est typée : ME pour une mention, RE pour une réponse à un message, et RT pour une ré-émission à l'identique (ou « retweet »). Le premier compte, @1, est dans une dynamique de spam : il promeut un site web auprès d'autres utilisateurs à sa portée. Ses cibles, @2 et @3, n'écrivent pas de messages, insensibles à son activité. Dans un second sujet, @4, @5 et @6 s'organisent une soirée cinéma et discutent de films. Enfin @7 et @8 préfèrent le théâtre et essayent -brièvement- de convaincre @5 de les rejoindre.

### 6.2.2 Les utilisateurs

Afin de compléter les informations et d'accroître la similarité avec celles provenant de Twitter, nous ajoutons la table 6.2, contenant les colonnes suivantes : *Jours* nombre de jours depuis la création du compte, M$essages$ nombre de messages émis, *Friends* nombre d'amis (abonnements), *Followers* nombre de followers. Par exemple, si A s'abonne aux contenus postés par B, alors A est le *follower* de B, et B fait partie des *Friends* de A. Twitter fournit aussi d'autres informations (photo de profil, de

TABLEAU 6.2 – Ensemble des données utilisateurs pour l'exemple illustratif *CinéTweets*

| Auteur | Jours | Messages | Friends | Followers |
|---|---|---|---|---|
| 1 | 10 | 1000 | 1500 | 15 |
| 2 | 1502 | 0 | 54 | 6 |
| 3 | 875 | 15 | 187 | 23 |
| 4 | 200 | 50 | 10 | 10 |
| 5 | 200 | 125 | 10 | 15 |
| 6 | 200 | 10 | 12 | 18 |
| 7 | 50 | 10 | 10 | 10 |
| 8 | 100 | 50 | 30 | 10 |

bannière, statut de vérification, nombre de comptes dans ses listes, nombre de tweets favoris, ...) dont l'inclusion rendrait l'exemple moins clair.

La table 6.2 se lit ainsi : l'utilisateur @1 a créé son compte il y a 10 jours, et il a déjà émis 1000 tweets. L'utilisateur @6 s'est abonné à 12 comptes. En retour, il a 18 followers / suiveurs.

Ces données ne montrent ici qu'une quinzaine de messages et une dizaine de comptes utilisateurs. Les quantités de messages échangés sur les réseaux sociaux sont bien supérieures, mais l'exemple fourni permet déjà, à la fois d'identifier rapidement qui sont les acteurs-clés du réseau, et de faire ressortir les difficultés.

Dans la suite de ce chapitre, nous ferons régulièrement référence à cet exemple illustratif. Nous espérons qu'il aidera à rendre notre propos plus clair.

## 6.3 Mesure de l'influence de l'utilisateur

Le premier axe du besoin consiste à élaborer un ou plusieurs scores d'influence. *L'influence* est souvent vue comme la capacité à susciter l'engagement, l'action de la part des autres. Elle est souvent confondue avec *l'expertise*, qui mesure la valeur de l'avis d'un utilisateur sur un sujet donné, et avec la *popularité*, mesurant le nombre de personnes connaissant un compte donné.

L'influence comme réaction d'autres utilisateurs doit prendre en compte l'influence propre des comptes qui réagissent. Un compte est plus influent s'il est proche de comptes eux-mêmes influents. L'expertise nécessite un travail de classification thématique des messages, afin de préciser à quel point un auteur se disperse, ou bien reste focalisé dans son domaine. Enfin, la popularité se résume à la taille de l'audience : le nombre de comptes ayant souscrit aux contenus publiés.

Pour calculer ces scores, de nombreuses informations sont déjà disponibles. Selon la plate-forme, elles sont parfois fournies en même temps que les messages. Dans cette section, nous les passons en revue et introduisons les caractéristiques supplémentaires à calculer.

### 6.3.1 Transformation d'indicateurs connus

Intuitivement, les acteurs-clés sont avant tout *actifs* sur le média social : une statistique simple les signalant repose sur la quantité de messages émis, que ce soit par jour, ou depuis la création du compte utilisateur. Cependant, sur les plate-formes numériques, de nombreux comptes sont automatisés, légitimement (par exemple, les journaux et sites d'information) ou non (notamment, le spam). L'influence ne consiste donc pas à être le plus actif : un niveau très bas de publication (par exemple, un tweet par semaine en moyenne) suffit à conférer une importance à un compte.

La mesure de l'impact d'un message passe, *a priori*, par la taille de son audience naturelle. Ceci explique l'engouement pour le nombre d'amis ou de suiveurs sur Facebook et Twitter : il s'agit d'une première mesure simple caractérisant un compte utilisateur à un instant $t$.

De même, sur Twitter, une autre statistique au niveau du message est disponible : le nombre de retweets (rediffusion à l'identique du message). En figure 6.2, nous montrons la répartition du

nombre d'utilisateurs (en ordonnée) selon le nombre de fois qu'ils ont été retweetés (en abscisse) ; cette information est calculée sur un corpus couvrant un mois de tweets émis, par et vers, 5 000 comptes. Les échelles sont logarithmiques : environ $10^5$ individus ont été retweetés 2 fois, alors qu'un seul individu a été retweeté 987 fois exactement [2].

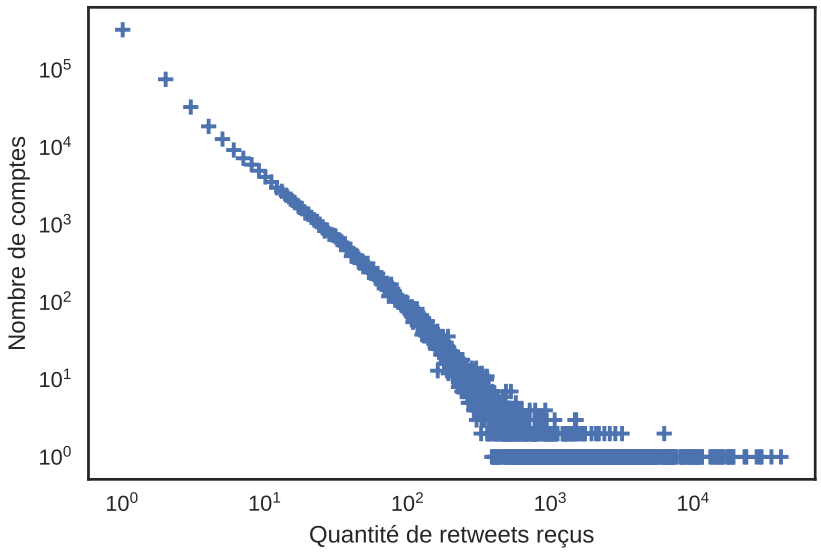



FIGURE 6.2 – Répartition du nombre de retweets

De cette figure, nous retenons qu'un retweet a plus de poids et d'importance lorsqu'il vise un « petit » compte, là où plusieurs centaines sont nécessaires pour faire une différence auprès de « grands » comptes. Autrement dit, le nombre de retweets (resp., de followers) n'est pas un score linéaire. Afin de mieux se représenter ce score, par exemple sur une échelle de 0 à 1, ou bien de 0 à 100, il faut utiliser un mécanisme de transformation.

Nous introduisons un paramètre, MAX_RT (resp., MAX_FOLLOWERS), correspondant au nombre maximal de retweets (resp., de followers) pour lequel le score maximal est attendu. Nous utilisons ensuite le logarithme en base MAX_RT (resp., MAX_FOLLOWERS) pour réaliser cette projection. Afin de fixer des valeurs pour ces deux paramètres, nous choisissons le nombre maximal constaté sur le réseau. Ainsi, en mai 2018, la personne la plus suivie sur le réseau Twitter est Katy Perry [3] avec plus de 100 millions de followers. Un paramètre MAX_FOLLOWERS = $10^8$ concède un score de popularité de (supérieur à...) 1 à Katy Perry, et permet de placer n'importe quel compte sur cette échelle de 0 à 1.

Ce mécanisme, décrit en équation 6.1, permet d'exprimer la taille de l'audience directe d'un compte par un score entre 0 et 1, que nous nommons **popularité**. Il respecte le principe, déjà vu dans d'autres scores [Rao et al., 2015], d'augmentation de la difficulté : il faut toujours plus de *followers* pour gagner le point de score suivant. Nous forçons la valeur minimale à 0, pour éviter l'infini négatif issu du logarithme : en conséquence, il n'y a pas de différence de score entre 0 et 1 *follower*.

$$popularite(u) = \max\left(0\ ;\ \log_{\text{MAX\_FOLLOWERS}}\left(\#Followers(u)\right)\right) \tag{6.1}$$

Notons que derrière l'action de l'abonnement, plusieurs vérités se cachent. Certains comptes, correspondant à des comptes réputés, sont ajoutés aux abonnements sans avoir besoin d'en lire un seul message, et peu d'interactions voire de lecture en découle. Par ailleurs, il existe un ajout suite à la lecture de plusieurs messages ayant retenu l'attention, résultant souvent en un lien plus fort ; il

2. ce qui explique le nombre de croix pour chaque valeur autour de $10^3$.
3. relevé sur https ://twittercounter.com/pages/100

demeure cependant trop difficile d'identifier automatiquement la différence d'intention dans l'action d'abonnement pour que nous la considérions dans nos travaux.

L'application aux retweets n'est pas aussi directe : il y a une dépendance à la manière de compter, et au domaine. Certains thèmes sont plus propices aux retweets de masse, alors que des enquêtes plus ciblées sur des domaines plus confidentiels n'ont pas d'intérêt en des valeurs trop élevées.

Nous avons décidé d'utiliser le mécanisme du retweet pour mesurer l'**expertise** d'un auteur : si un compte se focalise sur une thématique donnée, et que ses messages sont régulièrement repris par son audience, c'est qu'il est reconnu comme un expert sur cette thématique. Plusieurs méthodes sont en concurrence pour extraire le thème ou annoter un texte, et peuvent être utilisées de manière interchangeable ici. Pour obtenir un score, nous utilisons la notation suivante :

— $\theta$, l'un des labels de thèmes parmi un ensemble de thèmes $\Theta$

— $\text{MAX\_RT}_\theta$, un paramètre de nombre maximal de retweets, qui peut être adapté selon le problème

— $\#\text{RT}_\theta(u)$, le nombre de retweets de thème principal $\theta$ dont est gratifié l'utilisateur $u$

$$expertise(u) = \max\left(0\ ;\ \max_{\theta\in\Theta} \log_{\text{MAX\_RT}_\theta}\left(\#\text{RT}_\theta(u)\right)\right) \tag{6.2}$$

L'équation 6.2 introduit cette mesure d'expertise : parmi tous les thèmes ou domaines que l'auteur peut aborder, son meilleur domaine est retenu. Sur celui-ci, un score, dont la structure est similaire à la popularité, est calculé ; pour calibrer le score nous utilisons le nombre de retweets total atteint sur le domaine visé, qui représente la somme des réémissions de messages tombant dans le domaine de prédilection. Ce score répond à quelques pré-requis : il dépend du domaine, illustre l'impact réel des messages, et reste facile à appréhender.

Cependant, l'expertise ne mesure pas la régularité de l'impact de l'utilisateur sur un domaine : il peut suffire d'un tweet très retweeté pour devenir expert. Pourtant, elle favorise aussi la production régulière de contenu en prenant en compte le nombre total de retweets, c'est-à-dire en agrégeant tous les tweets thématiques émis par l'auteur dans le jeu de données.

### 6.3.2 L'influence comme position centrale dans le réseau

La littérature a introduit plusieurs scores dérivés de graphes : dans un graphe $G = (V, E)$, où l'ensemble des nœuds représente les utilisateurs, liés entre eux par des arcs $e \in E$ représentant soit l'interaction, soit la confiance, soit autre chose (souvent, des liens « d'amitié » ou de contact), plusieurs mesures sont calculables.

Nous pensons que les liens d'abonnement ou d'amitié sont trop faibles, c'est-à-dire qu'ils ne garantissent pas suffisamment de proximité sociale pour servir de support à nos mesures. Nous construisons alors un graphe d'interactions $G_I = (U, E_I)$, à partir de l'utilisation des mentions. Une mention est une référence à un autre compte utilisateur, faite dans le texte même du tweet. Dans G, U est l'ensemble des comptes utilisateurs, et $E_I$ sont les arcs dirigés de l'auteur de l'interaction vers sa destination, valués (ajout d'un poids unitaire par interaction ponctuelle).

Deux mesures en particulier sont présentées comme corrélées à l'influence : la centralité et le PageRank [Kwak et al., 2010], qui ont été présentées en chapitre 3. Afin d'obtenir rapidement un score d'**influence** thématique, nous nous basons sur ce PageRank calculé sur un graphe des mentions thématiques. Le texte des messages porteurs de mentions est analysé pour en extraire une thématique $\theta$, et ainsi attribuer une thématique à chaque interaction. Il en résulte $G_I^\theta$ le graphe des mentions de thème $\theta$ : $G_I^\theta = (U, E_I^\theta)$.

Le PageRank [Page et al., 1999] répond à l'intuition qu'un nœud référencé par des nœuds ayant déjà un bon score, voit son score amélioré. Nous considérons qu'il prend en compte la « qualité », l'influence des nœuds. Par construction, le Pagerank est calculé pour tout le graphe, et stocké dans

une table. En équations 6.3 et 6.4, le score d'influence correspond à la valeur du PageRank du nœud-utilisateur $u$, dans le graphe choisi (thématique, ou non).

$$influence_\theta(u) = \quad pagerank(u, G_I^\theta) \tag{6.3}$$

$$influence(u) = \quad pagerank(u, G_I) \tag{6.4}$$

Cet indicateur est facile et rapide à calculer, atout notable pour son éventuelle intégration dans un produit d'analyse de réseaux sociaux. De plus, son résultat peut être normalisé, répartissant les scores sur une échelle pratique à comprendre, entre 0 et 1 ou bien entre 0 et 100, par exemple.

La centralité d'un nœud $n$ est la somme de tous les plus courts chemins entre $n$ et tous les autres nœuds du graphe. Elle dépend donc finalement plus de la proximité d'un *hub*, que de la quantité de fois où un message a été effectivement diffusé par $n$ comptes dans le réseau. Dans un petit réseau social, le nœud central peut souvent être vu comme un acteur-clé : sa position lui confère une facilité à propager l'information. Cependant, dans un grand réseau, nous estimons que le Pagerank correspond mieux à la notion d'influence. En effet, il ne s'agit plus de toucher l'intégralité du réseau rapidement, mais d'être convaincant auprès d'une audience suffisamment large. Un nœud périphérique, de faible centralité, peut être régulièrement cité et pris pour référence par un grand nombre d'autres nœuds.

TABLEAU 6.3 – Scores d'influence

| Indicateur | Description |
|---|---|
| Nombre d'abonnés | Dépend de la population du réseau |
| **Popularité** | Paramétrable ; score sur 1 |
| Nombre de retweets | Ne prend pas en compte les thématiques |
| **Expertise** | Considère les thèmes abordés ; score sur 1 |
| Centralité | Dépend de l'ensemble du graphe |
| **Influence** | Classement ; dépend surtout de l'environnement proche |

Nous récapitulons ces indicateurs d'influence en table 6.3. Le cœur de notre contribution consiste d'une part à obtenir des scores aisés à exploiter, et d'autre part à utiliser conjointement l'influence, l'expertise et la popularité pour comparer l'impact des comptes utilisateurs. En effet, les simples nombres d'abonnés ou de réémission à l'identique ne sont finalement pas aisés à comparer, car la répartition de leurs valeurs dépend de la plate-forme analysée ainsi que de la diffusion d'une thématique donnée.

Nous proposons un exemple concret des mesures de l'influence, en les appliquant à l'exemple *CinéTweets*.

### 6.3.3 Application à l'exemple *CinéTweets*

La figure 6.3 illustre le graphe des interactions $G_I$ entre les comptes utilisateurs introduits dans l'exemple *CinéTweets*. Les flèches (portions d'arc en gras) sont dirigées des auteurs vers les comptes mentionnés (ou retweetés, ou répondus). La spatialisation et la longueur des arcs ne portent pas de signification particulière.

Sur ce graphe, nous pouvons calculer les scores d'*influence*, que nous montrons en table 6.4. Grâce à leurs échanges répétés, les utilisateurs @4, @5 et @6 obtiennent les meilleurs scores. Notamment, @5 bénéficie de retweets et obtient le meilleur score d'influence, avec 30 points. Pour clarifier la lecture, nous montrons ces scores en points, sur 100.

Malgré ses efforts répétés, le compte de spam, @1, n'ayant bénéficié d'aucun retweet ou mention, se distingue ici par ses scores très faibles. Avec 23 followers, correspondant à un score de popularité de 17, le compte le plus populaire est @3.

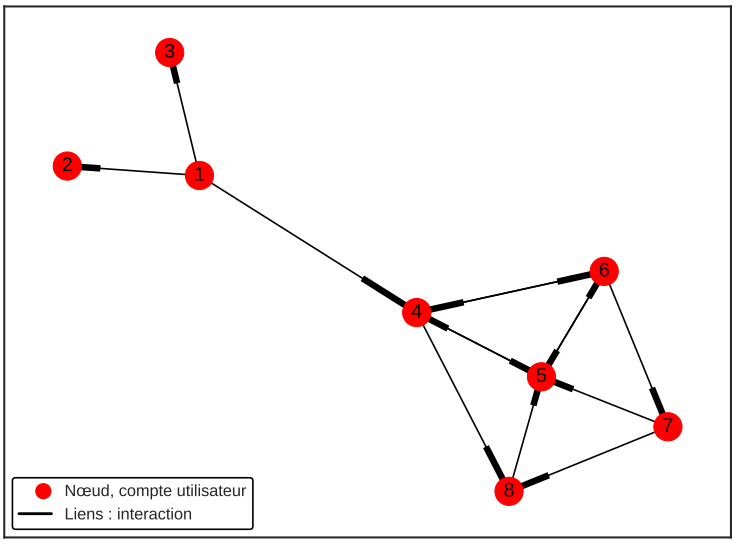



FIGURE 6.3 – Visualisation du graphe des interactions dans l'exemple *CinéTweets*

Pour mesurer le score d'expertise, nous fixons le paramètre MAX_RT = 10 : ce score n'est pas atteint dans l'exemple minimal, mais est raisonnablement bas pour gratifier les comptes retweetés. Dans une application réelle, sur un corpus de taille habituelle, ce paramètre peut correspondre à la valeur maximale constatée dans le graphe des retweets.

@4 a été plus retweeté sur le cinéma que @5, et passe pour un expert en ce domaine. Sur le théâtre par contre, @5 bénéficie d'un score d'expertise de 47.

TABLEAU 6.4 – Scores d'influence appliqués à *CinéTweets*

| Compte | Followers | Popularité | Retweets | Expertise | Mentions | Influence |
|---|---|---|---|---|---|---|
| 1 | 15 | 14 | 0 | 0 | 0 | 3 |
| 2 | 6 | 9 | 0 | 0 | 1 | 3 |
| 3 | 23 | **17** | 0 | 0 | 1 | 3 |
| 4 | 10 | 12 | $2_{cinema}$ | $30_{cinema}$ | 3 | 18 |
| 5 | 15 | 14 | $1_{cinema}$, $3_{theatre}$ | $47_{theatre}$ | 4 | **30** |
| 6 | 18 | 15 | 0 | 0 | 2 | 23 |
| 7 | 10 | 12 | 0 | 0 | 1 | 9 |
| 8 | 10 | 12 | 0 | 0 | 2 | 10 |

### 6.3.4 Discussion sur l'influence

Par abus de langage, la détection d'influenceurs recouvre une énorme variété de méthodes et de réalités. Tant la définition d'influence que le périmètre de recherche (période temporelle, corpus thématique, géographique, corpus de documents) sont variables au gré des fonctionnalités réclamées pour l'analyse des réseaux sociaux.

Les trois indicateurs que nous avons introduits, **influence**, **expertise** et **popularité**, permettent de répondre à une partie des besoins soulevés et facilitent la lecture des données : difficile de savoir, *a priori*, combien de retweets ont été émis dans une thématique donnée et donc quel score attendre des experts du domaine. La construction de ces indicateurs permet de s'affranchir de cette connaissance

trop spécifique et trop peu répandue, pour obtenir des scores clairs et réels. Le score d'*influence* lui-même n'est certes pas nouveau, mais il est indispensable pour caractériser la valeur de la position sociale occupée par le compte utilisateur dans le graphe des interactions.

## 6.4 Caractérisation du comportement par le profil-type

Le second volet de ce chapitre, portant sur le profilage à partir des comportements d'une population d'utilisateurs, est plus complexe. Il s'agit de donner les clés de compréhension sur le comportement habituel d'un compte utilisateur, tout en restant assez flexible pour ne pas se focaliser uniquement sur une mesure brute de son activité.

### 6.4.1 Construction de Profils

Nous définissons le *profil* $\pi_u$ d'un compte comme l'ensemble des caractéristiques, numériques ou non, décrivant le compte et son comportement. Il permet de conserver les modifications de ces caractéristiques dans le temps et de calculer des similarités avec d'autres comptes. Notamment, nous introduisons la notion de *profils-types*, notés $p$, qui permettent de catégoriser les utilisateurs, mode de représentation plus compréhensible qu'une longue liste de caractéristiques numériques.

Pour amener des éléments de réponse, nous avons construit un modèle de représentation de l'utilisateur le long de cinq **aspects** du comportement, ce qui structure le problème et en facilite la compréhension. Ce découpage permet de se focaliser sur l'aspect d'intérêt pour le métier (intelligence économique, journalisme, renseignement...), qui peut évoluer selon les besoins, selon les comportements-types recherchés. De même, ce découpage peut être appliqué à d'autres réseaux sociaux, même s'il est conçu pour coller aux médias sociaux tels que Twitter. Par exemple, pour s'adapter à Facebook, il faudrait compter de manière distincte les « j'aime » par émoticônes.

Le premier aspect, Biographie, regroupe tous les éléments parlant de l'identité ou des paramètres de préférence de l'utilisateur. Cet aspect est un peu particulier puisqu'il contient de nombreuses caractéristiques non numériques.

Le second aspect, Style, concerne le style d'écriture des messages publiés. En effet, il est simple et globalement informatif de mesurer la taille des messages, l'utilisation de ponctuation, d'émoticônes, etc.

Le troisième aspect, Media, adresse la variété de contenus et d'actions possibles. Il regroupe les comptages d'émission de liens internet, photos, vidéos, publications de messages, de commentaires, etc.

Le quatrième aspect, Interaction, mesure et type les actions impliquant d'autres comptes utilisateur sur le réseau. Il s'agit de l'aspect le plus social.

Le cinquième aspect, Temporel, regroupe la représentation en série temporelle des actions accomplies par le compte utilisateur. Il contient la dimension dynamique du comportement.

Nous explicitons plus précisément cette contribution en listant les caractéristiques mesurées, aspect par aspect, dans les paragraphes suivants.

#### 6.4.1.1 Biographie

Cet aspect regroupe les caractéristiques permettant de se faire une idée de l'identité de la personne. Des calculs spécifiques de comparaison de pseudonymes par exemple, peuvent faire émerger des ressemblances, des motifs récurrents [4]. Des fonctionnalités supplémentaires, hors du cadre des travaux de cette thèse, permettraient de transformer le champ « *location* » en coordonnées GPS permettant l'affichage sur une carte d'un ensemble d'utilisateurs, par exemple. À chaque message reçu, nous obtenons et conservons les caractéristiques suivantes :

4. Souvent, des botnets utilisent une fonction pour générer des noms, par exemple PRENOM+chiffre, l'utilisation trop fréquente de signes comme l'underscore _, etc.

— $name_u$ le nom complet du compte utilisateur,
— $screen-name_u$ le pseudonyme court donné par l'utilisateur,
— $\mathrm{ID}_u$ l'identifiant unique du compte,
— $biography_u$ un champ libre de texte, un résumé de l'identité,
— $location_u$ un champ libre de texte, permettant d'indiquer une localisation,
— $creation-date_u$ l'instant de création du compte utilisateur,
— $timezone_u$ le fuseau horaire d'affichage,
— $language_u$ la langue de l'interface,
— $source_u$, c'est-à-dire l'application et/ou la plateforme physique utilisée pour envoyer un message, par exemple type de téléphone,
— $followers-count_u$ le nombre d'abonnés,
— $friends-count_u$ le nombre d'abonnements à d'autres comptes,
— $statuses-count_u$ le nombre de messages émis.

#### 6.4.1.2 Style

Cet aspect regroupe des comptages variés : nombre de caractères, quantité de ponctuation, de hashtags, etc. En effet, la manière de rédiger un message diffère significativement entre comptes, certains montrant des habitudes bien ancrées (par exemple, le hashtag en fin de message). Les caractéristiques sont d'abord calculées pour chaque texte $t$ ; puis une seconde étape, consiste à calculer les scores moyens d'un utilisateur $u$, à partir des caractéristiques mesurées pour chacune de ses publications dans le jeu de données.

— $l_u$ la longueur moyenne du message,
— $pc_u$ la quantité moyenne de ponctuation (telle que ! ?.),
— $emoj_u$ la quantité moyenne d'emojis (caractères Unicode),
— $emo_u$ la quantité moyenne d'émoticônes classiques parmi [ :), :-), :D, :(, :/],
— $hb_u$ le nombre moyen de hashtags en début de texte,
— $he_u$ le nombre moyen de hashtags en fin de texte,
— $hm_u$ le ratio moyen du nombre de hashtags/mots,
— $s_u$ sentiment moyen (polarité $\in [-1;1]$),
— $\theta_u$ la thématique la plus fréquente (un label est calculé par un détecteur de thématique pour chacun des textes).

#### 6.4.1.3 Média

Chacun des objets sociaux $o \in \Omega$, défini en équation 6.5, contenu dans un message est un indice d'un *comportement média*. En effet, nous voyons ces éléments comme des contenus voués au partage, à la diffusion en tant qu'objets.

$$\Omega = [media, \mathrm{URL}, hashtag] \tag{6.5}$$

Ainsi, pour chaque type d'objet social $o$ nous calculons $o^d$ son nombre moyen d'occurrence quotidienne, $o^r$ son nombre moyen d'occurrence par publication, et $o^q$ sa quantité totale, comme introduit en équation 6.6. Avec $\mathrm{T}_d$ l'ensemble des messages publics émis par l'utilisateur $u$ durant le jour $d \in \mathrm{D}$ (D la période d'analyse globale), et pour chaque texte $t$ nous notons $o^t$ le nombre d'objets sociaux de type $o$ présents :

$$o^d = \frac{1}{|\mathrm{D}|} \sum_{day\ d} \sum_{t \in \mathrm{T}_d} o^t \tag{6.6}$$

$$o^q = \sum o^t \tag{6.7}$$

$$o^r = \frac{o^q}{|\mathrm{T_D}|} \tag{6.8}$$

Ces trois indicateurs de la même modalité, l'activité sur l'émission d'objets sociaux, permettent de faire la différence entre l'émission de nombreux objets dans un unique message (par exemple, lorsque tous les mots sont des hashtags), la variation journalière de cet usage, et éventuellement d'un comportement où il y aurait des textes « avec » et « sans » objets sociaux, signe de la présence de deux émetteurs différents derrière le même compte utilisateur. Nous modélisons donc cet aspect par les caractéristiques suivantes :

— Pour les médias : $media_u^d$, $media_u^q$, $media_u^r$

— Pour les URLs : $url_u^d$, $url_u^q$, $url_u^r$

— Pour les hashtags : $hashtag_u^d$, $hashtag_u^q$, $hashtag_u^r$

#### 6.4.1.4 Interaction

De la littérature, nous réutilisons un ratio $FrFo_u = \frac{|\mathrm{F}riends_u|}{1+|\mathrm{F}ollowers_u|}$ [Varol et al., 2014]. L'ajout de 1 au nombre de *followers* (abonnés), en dénominateur, permet d'éviter la division par zéro. Ce ratio, indépendant des valeurs absolues, permet de repérer les comptes écouteurs (nombreux abonnements, faible nombre de *followers*) des comptes émetteurs ou connus (nombreux *followers* par rapport aux abonnements : l'utilisateur n'est pas là pour lire).

Nous calculons $TJours_u$, le ratio du nombre de textes émis par jour, représentant l'activité moyenne globale (*statuses-count* divisée par le nombre de jours depuis la création du compte). En effet, quelques comptes se distinguent par leur régularité sans faille, parfois autour de 200 messages par jour.

De plus, nous introduisons de nouveaux indicateurs en équation 6.9 : $sociability_u$ compte le nombre total de mentions dans un ensemble de messages de même auteur, et $diversity_u$ est lié à |U| le nombre d'utilisateurs *différents* mentionnés dans un ensemble de messages (typiquement, sur une journée). Un mécanisme pour attirer l'attention consiste à mentionner n'importe qui, suscitant l'envoi d'une notification et possiblement, un *follower* supplémentaire.

$$originality_u = \frac{|\mathrm{RT}|}{|\mathrm{T}|} \tag{6.9}$$

$$diversity_u = \frac{|\mathrm{U}|}{|\mathrm{T}|} \tag{6.10}$$

Enfin, $originality_u$ est le ratio de messages originaux (excluant les réémissions à l'identique, de type *retweet*). De nombreux humains se contentent de retransmettre des messages pour marquer leur approbation à un message donné ; d'autre part, certains robots utilisent ce mécanisme pour améliorer la diffusion de contenus provenant de leur faction, notamment dans le cadre d'un botnet.

#### 6.4.1.5 Temporel

L'aspect temporel permet de bien représenter les périodes d'activité d'un compte donné. À partir de l'ensemble des dates d'émission des messages, nous calculons un histogramme $hist_u$ représentant l'activité d'un compte, heure par heure sur une semaine. Cet histogramme permet de conserver un comportement moyen et éventuellement d'observer des récurrences et/ou anomalies.

Cependant, la figure 6.4 montre la quantité de messages postés par deux utilisateurs différents pendant une même période d'une heure. À gauche, l'utilisateur semble actif pendant une dizaine de minutes : en quelques instants, il publie et retransmet des messages, jusqu'à 12 actions en une minute, puis se déconnecte. À droite, les niveaux d'activités sont plus faibles mais nettement plus réguliers.

L'histogramme s'interrompt à 13h à cause de la donnée sélectionnée (par tranches d'une heure) : ce compte a continué au même rythme les heures suivantes. Ainsi, le nombre de messages par heure ne suffit pas à synthétiser un comportement.

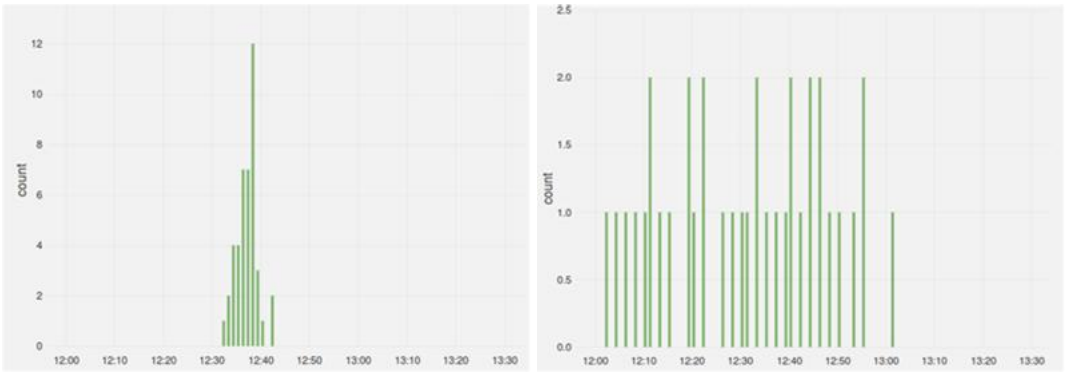



FIGURE 6.4 – Répartition temporelle des actions de deux utilisateurs

Pour pallier ce manque, le rythme peut faire une différence entre un automate simpliste, qui émet toutes les $n$ minutes, et un utilisateur régulier. Inspiré des statistiques sur les écarts temporels entre des événements extraits de communautés de Q&A [Fu et al., 2016], nous calculons un temps moyen entre deux messages consécutifs $\mu_u$ et son écart-type $\sigma_u$ pour chaque compte utilisateur $u$.

Trois éléments sont donc conservés ici : la table $hist_u$, ainsi que les réels $\mu_u$ et $\sigma_u$.

### 6.4.2 Profils-types

L'élaboration de profils-types a pour objectif de permettre de rapidement positionner un individu en résumant très succinctement les dizaines de valeurs calculées stockées dans un profil.

#### 6.4.2.1 Définitions

À partir des profils $\pi_u$ définis pour tout utilisateur $u \in U$, nous souhaitons extraire une répartition le long d'un nombre réduit de catégories.

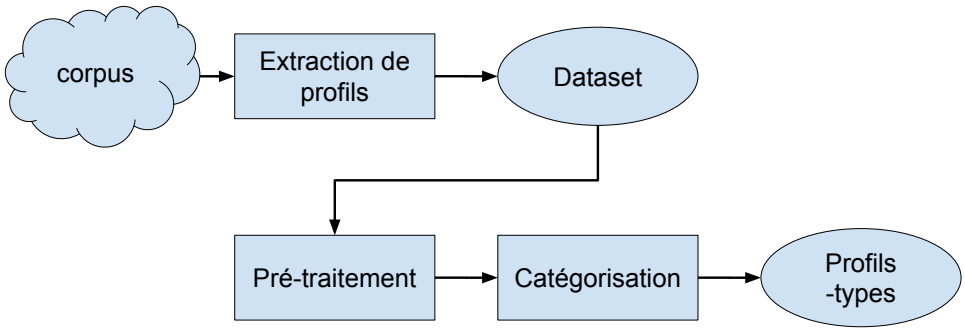



FIGURE 6.5 – Classification des profils et obtention de profils-types

La figure 6.5 montre la chaîne de traitement pour la détermination des profils-types. Un jeu de données de profils (« dataset ») est extrait d'un corpus de messages échangés. Après un éventuel pré-traitement (ou nettoyage), ce jeu de données est ingéré par un algorithme de catégorisation. Il en résulte l'attribution de chacun des profils $\pi_u$ à un label de profil-type $\Pi_i$. L'algorithme doit être capable de calculer le profil-type d'un nouveau compte sans procéder à une remise à plat systématique de sa taxonomie.

Selon le modèle du type des utilisateurs fourni, la caractérisation du comportement peut prendre trois formes :

— un modèle expert. Il s'agit d'un découpage selon des caractéristiques pré-établies. Les valeurs de profils-types $\Pi_1, \Pi_2, ...\Pi_k$ sont fournies directement. Le modèle n'a plus qu'à associer chaque profil $\pi_u$ avec le profil-type adéquat.

— classification supervisée grâce à un ensemble de données annotées. Par exemple, quelques comptes utilisateurs sont identifiés comme représentatifs de chacun des profils-types, et les concrétisent. Ces données alimentent ensuite un classifieur supervisé.

— classification non-supervisée, dans le cas où une catégorisation n'est pas fournie *a priori*. La population de profils $\pi_u \ \forall \ u \in \mathrm{U}$ alimente alors un algorithme de *clustering*, tel un k-moyennes, pour en extraire des profils-types $\Pi_1, \Pi_2, ...\Pi_k$ adaptés au jeu de données. Selon la méthode choisie, la valeur de $k$ peut être choisie ou déterminée par l'algorithme.

Nous ne prévoyons pas de disposer d'un expert, ni d'un jeu de données annotées. L'approche de classification non-supervisée est donc retenue.

#### 6.4.2.2 Traitement des données

Tout d'abord, un filtrage des caractéristiques doit être effectué : selon le besoin, la catégorisation est calculée le long d'un ou plusieurs aspects, à l'exclusion d'autres. De plus, certaines caractéristiques doivent être remplacées par leur logarithme, à cause de leur dispersion initiale. C'est le cas pour les quantités de contenus média publiés (aspect *Média*), pour les ratios $FrFo_u$ et $TJour_u$, et pour les caractéristiques temporelles $\mu_u$ et $\sigma_u$.

Dans une seconde phase, les données doivent être normalisées pour obtenir une distribution centrée en 0, d'écart-type 1, souhaitable pour certaines méthodes d'apprentissage statistique comme les k-moyennes. Après quoi un autre traitement, de réduction de dimensionnalité, est requis pour éviter la « malédiction de la dimensionnalité » des méthodes basées sur la distance euclidienne, réputée souffrir au-delà de 10 dimensions : nous retenons l'ACP [5] pour réduire la dimension du problème.

Enfin, de façon adaptée aux données d'entrée, un algorithme de classification ingère le jeu de données et produit les labels $\Pi_1, \Pi_2, ...\Pi_k$ de profils-types.

#### 6.4.2.3 Qualité attendue et méthodes d'évaluation

Avec une approche supervisée, nous disposons soit des profils-types, *a priori*, soit d'une annotation des comptes utilisateurs pour permettre l'entraînement d'un algorithme d'apprentissage supervisé. Le premier cas est propice à une validation empirique, bien que limitée. Dans le second cas, une précision et un rappel peuvent être calculés, ainsi qu'une fiabilité. La matrice de confusion peut aider à comprendre d'où provient l'erreur éventuelle.

Avec une approche non-supervisée, il existe des mesures telle que la mesure de Calinski-Harabaz [Caliński and Harabasz, 1974], qui est un ratio entre dispersions inter- et intra- clusters. Une valeur élevée signifie des clusters mieux formés, plus éloignés les uns des autres. Cela permet de choisir une valeur de $k$ adaptée à la distribution des données. Il peut être utile de chercher un compromis pour que cette valeur reste faible : inutile d'obtenir des dizaines de profils-types, l'objectif est de simplifier la vue des comportements de la population de comptes utilisateurs.

Plus précisément, le score $s(k)$ d'un découpage en $k$ clusters est donné en équation 6.11. $\mathrm{W}_k$ représente la dispersion interne (*within*), $\mathrm{B}_k$ la dispersion inter-clusters (*between*). Pour ces deux matrices, seule la trace ($\mathrm{Tr}(\cdot)$, somme des valeurs de la diagonale) compte. N est la quantité d'observations dans le jeu de données, $\mathrm{C}_q$ l'ensemble des observations dans le cluster $q$, $c_q$ le centre du cluster d'indice $q$, $c$ le centre du jeu de données.

5. Analyse en Composantes Principales

$$s(k) = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N-k}{k-1} \tag{6.11}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{6.12}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c)(c_q - c)^T \tag{6.13}$$

Bien sûr, cette analyse non supervisée n'échappe pas aux risques associés : les profils-types peuvent ne pas correspondre à l'application recherchée, et manquer de sens. Cependant, le découpage des dimensions par aspects donne la possibilité de calculer des aspects-types, qui, plus près des données, pourront mieux caractériser un volet comportemental plus précis.

#### 6.4.2.4 Application à *CinéTweets*

Une analyse de notre exemple-jouet peut clarifier le fonctionnement de notre modèle. Dans un premier temps, nous allons construire les profils $\pi_u$ simplifiés correspondant à nos données. Les caractéristiques retenues sont les suivantes :

- Biographie
  - *jours*, nombre de jours depuis la création du compte ;
  - *messages*, nombre de messages émis ;
  - *friends*, nombre d'abonnements ;
  - *followers*, nombre d'abonnés ;
- Style
  - *emojis*, nombre d'émoticônes utilisées ;
- Média
  - *media*, le nombre d'objets sociaux émis ;
- Interaction
  - *frfo*, le ratio du nombre d'abonnements sur le nombre d'abonnés ;
  - *actions*, nombre d'actions sociales effectuées (de mentions émises) ;
- Temps
  - *tj*, tweets par jour ;

Afin de ne pas forcer nos observations à correspondre à une typologie pré-établie, dont nous ne disposons pas, nous suivons une approche non-supervisée.

Sur ces profils, nous appliquons un rééchelonnement pour limiter le poids relatif des variables ; puis nous utilisons l'ACP pour réduire la dimension du problème. Nous conservons ici 86% de la variance en limitant le nombre de dimensions à 3. La figure 6.6 montre le poids de chaque variable pour chacun des nouveaux axes issus de l'ACP, en reprenant chacune des variables des profils $\pi_u$.

Enfin, nous lançons un k-moyennes sur la projection de notre exemple. Étant donnée la taille très réduite du jeu de données, l'indicateur de Calinksi-Harabaz recommande d'utiliser $k = 8$, soit le nombre d'observations. À la place, nous fixons $k = 3$, comme compromis entre la diversité des profils et la taille de la population. La table 6.5 montre les labels de profils-types pour notre petite population. Le compte @1 est placé dans son propre cluster $p_2$ : beaucoup d'actions, de publications et d'objets (le lien qu'il ne cesse d'émettre) l'isolent du reste de la population. @2 et @3 n'ont rien émis, ce qui les distingue. Enfin, le groupe actif sur le cinéma et théâtre avaient été construits autour des mêmes valeurs de nombre d'amis, d'abonnements, etc. Sans surprise, ils sont regroupés au sein du même profil-type, $p_1$.

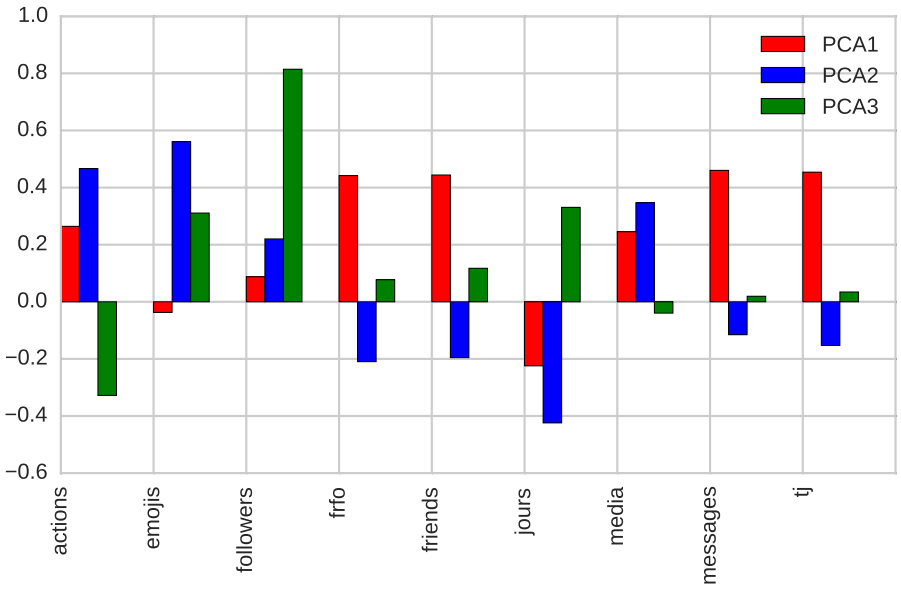



FIGURE 6.6 – Visualisation du poids des caractéristiques dans les 3 axes de l'ACP

TABLEAU 6.5 – table des étiquetages par profil-type

| Compte | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Profil-type | 2 | 0 | 0 | 1 | 1 | 1 | 1 | 1 |

## 6.5 Caractérisation de la position par le rôle

Troisième pilier de notre approche de caractérisation de l'utilisateur, l'analyse du rôle de l'utilisateur vise à établir une typologie des positions dans le graphe. Le rôle $r$ permet lui aussi de catégoriser l'utilisateur, non pas à partir des caractéristiques numériques issues de ses publications, mais à partir de sa position dans un graphe social. Ces rôles permettent ainsi, dans un second temps, de mieux caractériser les communautés qui seront découvertes sur le même graphe social.

### 6.5.1 Un algorithme de détection de rôles-types : RolX

De la même façon que les profils-types étiquettent les comptes utilisateur selon leur comportement, nous recherchons un étiquetage en fonction des positions-types occupées dans le graphe social. En effet, ces positions ne dépendent pas uniquement des quantités d'interaction mesurées précédemment, et sont informatives de la fonction accomplie par le compte : centre d'une communauté, interface entre groupes distants, récepteur isolé... autant de qualificatifs qu'un système automatique pourrait attribuer.

Parmi les méthodes revues en chapitre 3, nous avons retenu RolX [Henderson et al., 2012] comme outil de détection de rôles. RolX calcule une matrice de caractéristiques issues d'un graphe, composée de mesures topologiques pour chaque nœud. Parmi ces mesures se trouvent les degrés entrant, sortant, la centralité, et des mesures basées sur les ego-network : le réseau proche du nœud central. Cette technique vise à rendre calculables les mesures gourmandes, dont notamment l'intermédiarité qui nécessite le calcul de tous les plus courts chemins du graphe.

$$\mathrm{H\,K} \approx \mathrm{V} \tag{6.14}$$

Après une étape de sélection de caractéristiques, regroupées en la matrice V (de taille $n \times f$, $n$ nœuds, $f$ caractéristiques), la méthode réalise une factorisation non-négative de matrice, qui aboutit en un clustering. Le résultat, illustré par l'équation 6.14, est composé de H la matrice nœud-rôle (de

taille $n \times r$, $n$ nœuds, $r$ rôles), et de K la matrice rôle-caractéristique ($r \times f$, $r$ rôles, $f$ caractéristiques), qui décrit chaque rôle $r_1 ... r_r$ avec des mesures topologiques. Appliquée à un graphe d'interaction entre utilisateurs, cette méthode décrit des comportements typiques.

### 6.5.2 Expérience sur l'exemple *CinéTweets*

Nous avons procédé au calcul des rôles à partir du graphe d'interaction $G_I$, lui-même construit à partir du corpus *CinéTweets*, à la recherche de 4 rôles-types comme suggéré dans la littérature [Henderson et al., 2012]. Tous les rôles ne sont pas équitablement répartis dans la population des comptes utilisateurs : un rôle n'est représenté que par un nœud.

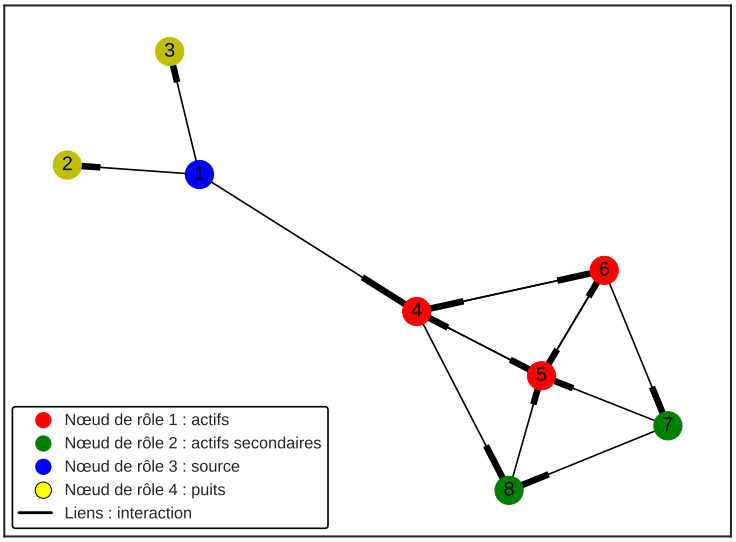



FIGURE 6.7 – Visualisation des rôles-types dans le graphe des interactions issu de *CinéTweets*

La figure 6.7 permet de visualiser de nouveau le réseau social introduit en exemple, cette-fois-ci coloré avec les rôles obtenus par l'approche non-supervisée *RolX*. La spatialisation et la longueur des arcs ne portent pas de signification particulière. Le nœud 1 occupe une position de source (rôle 3) : des arcs en sortent, sans entrée. À l'opposé, les nœuds 2 et 3 sont vus comme des puits (rôle 4). Les comptes 4, 5 et 6 sont les plus actifs, à la fois émetteurs et récepteurs ; de plus ils occupent une position centrale dans ce graphe (rôle 1). Enfin, les nœuds 7 et 8 ont un rôle d'actifs secondaires (rôle 2) : ils ont plus reçu qu'émis des interactions.

Notons que ces qualificatifs sont une interprétation de ces rôles à l'aide de la figure : les labels ne résultent pas d'un modèle initial, mais bien de notre analyse. Cette lecture est accompagnée de données concernant la répartition de quelques caractéristiques (degré entrant, degré sortant, pagerank) pour chaque rôle-type, ainsi que de visualisations de parties du graphe (lorsque le nombre de nœuds le permet). Ainsi, le choix de 4 rôles-types correspond plus à la faisabilité d'une interprétation humaine, qu'à une réalité du jeu de données.

### 6.5.3 Discussion sur le rôle

Initialement, suivant l'esprit des travaux sur le découpage Agent-Groupe-Rôle [Ferber and Gutknecht, 1998], le rôle représente la contribution d'un compte utilisateur à un groupe. Les agents peuvent

changer de rôle au cours du temps, et peuvent afficher plusieurs rôles au même moment, leur donnant des capacités et responsabilités.

Transposé à l'analyse des réseaux sociaux numérique, ce modèle imposerait de découper les différentes actions de l'utilisateur selon son ou ses appartenances à des groupes sociaux, ce qui n'a rien de trivial pour des arcs sans interaction. De plus, la détection de communautés autorisant la multi-appartenance reste un problème mal défini, difficile, malgré l'apparition récente d'algorithmes prometteurs.

Loin de cet idéal du « rôle de l'utilisateur dans le groupe », le rôle n'est ici qu'un label attribué aux nœuds d'un graphe, une catégorisation dont la taxonomie évolue selon l'instance du graphe donnée en entrée.

## 6.6 Synthèse

La détection des acteurs-clés d'un réseau est une tâche plus complexe qu'il n'y paraît. S'il est commun de mettre en avant des scores d'influence ou de capital social, la diversité des besoins est plus large et ne se satisfait pas d'un unique indicateur. Cependant, il ne faut pas sombrer dans l'excès de caractéristiques : l'objectif est de comprendre qui sont les auteurs des textes des messages collectés, en caractérisant les comportements et positions sociales.

Dans ce chapitre, nous avons présenté et structuré plusieurs mesures : influence, expertise, popularité. Elles répondent au besoin d'analyse et fournissent des scores clairs. Pour aller plus en profondeur et tirer profit de la richesse des données à disposition, nous avons organisé un modèle de représentation de l'utilisateur, que ce soit de ses données d'identité, de comportement, ou de positionnement dans un graphe social. Trop prolifiques, nous résumons ces données grâce au calcul de profils-types, établissant de grandes catégories du comportement, et par le calcul de rôles-types, qui éclaircit beaucoup la dimension sociale. Nous avons illustré ces méthodes en exploitant un jeu de données artificielles, *CinéTweets*, fourni en exemple. Sur ce jeu, nous avons calculé les scores d'influence, profils, profils-types, rôles et rôles-types, illustrant les méthodes introduites.

Le modèle de données, les fonctionnalités associées et la méthode d'utilisation constituent l'un des pans de travaux ayant donné lieu à un dépôt de brevet auprès de l'Office Européen du Brevet en octobre 2017.

Au-delà des comptes utilisateurs, les réseaux sociaux voient émerger des comportements de groupe, dont l'impact sur la propagation des messages et opinions est important. Le chapitre suivant s'intéresse à la détection de ces groupes d'utilisateurs, et à leur caractérisation.

CHAPITRE 7

# Détection et caractérisation de groupes d'utilisateurs

Nous nous penchons dans ce chapitre sur un phénomène social complexe : la présence de communautés émergentes dans les réseaux sociaux. Nous pensons qu'un groupe d'utilisateurs, individuellement peu influents, peut constituer en tant que groupe un acteur influent pour le reste du réseau s'il est concentré sur un thème donné. Il mérite alors d'être identifié en tant que *communauté thématique*. Dans le chapitre 4, nous avons observé différentes modélisations d'un réseau social par un ou plusieurs graphes ; nous avons aussi présenté les principaux algorithmes de détection de communautés ou de couvertures. Très peu de travaux se penchent à la fois sur les interactions entre utilisateurs, et sur l'exploitation du texte de leurs échanges. Ainsi, les communautés trouvées portent trop souvent exclusivement soit sur l'interaction, soit sur les centres d'intérêt des comptes. Or, nous sommes à la recherche de groupes présentant à la fois une forte interaction ainsi qu'un centrage fort sur une thématique, ces deux conditions étant propices pour considérer le groupe en tant qu'acteur d'influence.

Dans la suite de ce chapitre, la section 7.1 précise notre objectif de recherche et décrit l'approche suivie ; la section 7.2 introduit le modèle retenu pour représenter et manipuler les relations et centres d'intérêt des comptes utilisateurs, et expose les outils nécessaires pour détecter et mesurer les communautés, et extraire les thématiques de textes. Ensuite, nous introduisons des mesures de cohésion de groupes en section 7.3. En effet, l'état de l'art s'intéresse surtout à des mesures topologiques, que nous complétons avec notre contribution : des mesures basées sur les thématiques évoquées dans les messages. La section 7.4 décrit un réseau social en exemple, *CinéTweetsÉtendu*, pour illustrer notre contribution. Enfin, la section 7.5 propose quelques éléments de discussion. Une synthèse conclut ce chapitre.

## 7.1 Analyse de structures sociales émergentes

Sans forcément inclure de fonctionnalité explicitement nommée « groupe » ou « communauté », les réseaux sociaux numériques servent de support à l'émergence de structures sociales, rassemblant des comptes utilisateurs par des liens d'interaction intenses. Le contour de ces communautés est flou ; plusieurs intuitions et définitions rivalisent dans la littérature. Comme nous l'avons signalé en chapitre 4, la majorité de la littérature sur le sujet soit se focalise sur la détection de communautés dans des graphes, sans considérer les textes des messages échangés, soit se concentre exclusivement sur les textes des messages sans prendre en compte les liens sociaux ou les interactions entre les comptes utilisateur.

Pour notre part, nous pensons les communautés présentes sur un réseau social comme construites par l'interaction entre leurs membres. Il ne s'agit donc pas simplement d'afficher une similarité ou de revendiquer une prise d'intérêt : un abonnement ne suffit pas pour intégrer un groupe. La publication

de messages impliquant les autres membres nous semble constituer un lien plus concret, plus fort et plus tangible.

Parmi ces communautés, toutes ne sont pas aussi fortement liées. En particulier, nous sommes à la recherche de groupes visibles comme des acteurs médiatiques : pour qu'ils aient un impact ou une influence, il faut qu'ils s'expriment de façon cohérente, si ce n'est coordonnée. Parmi l'ensemble des communautés visibles sur le réseau social, nous cherchons un type particulier de communautés : des *communautés thématiques*, notion que nous clarifions en définition 7.1.1.

**Definition 7.1.1.** *Communautés thématiques :* il s'agit d'un ensemble de comptes utilisateurs, satisfaisant ces deux conditions :

— les membres sont densément connectés. Ils interagissent, discutent entre eux et partagent des contenus ;

— les membres sont intéressés par les mêmes thématiques et partagent des centres d'intérêt.

Afin de trouver ces communautés thématiques, nous proposons une méthode pour détecter puis caractériser les communautés actives sur un réseau social. La figure 7.1 montre les étapes de cette méthode : à partir d'un corpus de messages, par exemple des tweets, qui contiennent des informations à la fois sociales et textuelles, nous extrayons des relations entre les comptes utilisateurs.

Ces relations reposent sur des fonctionnalités variées proposées par la plate-forme, et nécessitent plusieurs graphes pour représenter les différents types d'interaction entre comptes utilisateurs. Sur ces graphes, un algorithme de détection de communautés est appliqué, résultant en un ensemble de groupes d'utilisateurs, rassemblés par leurs interactions. Des mesures topologiques évaluent la force de liaison interne de chacune de ces communautés, fournissant des informations à l'étape finale, d'analyse des communautés.

Parallèlement, une analyse sémantique de chacun des messages en extrait des thématiques, représentatives d'ensembles de documents, pour caractériser les centres d'intérêt des auteurs, membres de groupes. Cet élément ouvre la voie à des « mesures thématiques » de la qualité des communautés, basées sur la similarité des textes émis par leurs membres. Ces mesures viennent enrichir l'analyse des communautés présentes sur le réseau social.

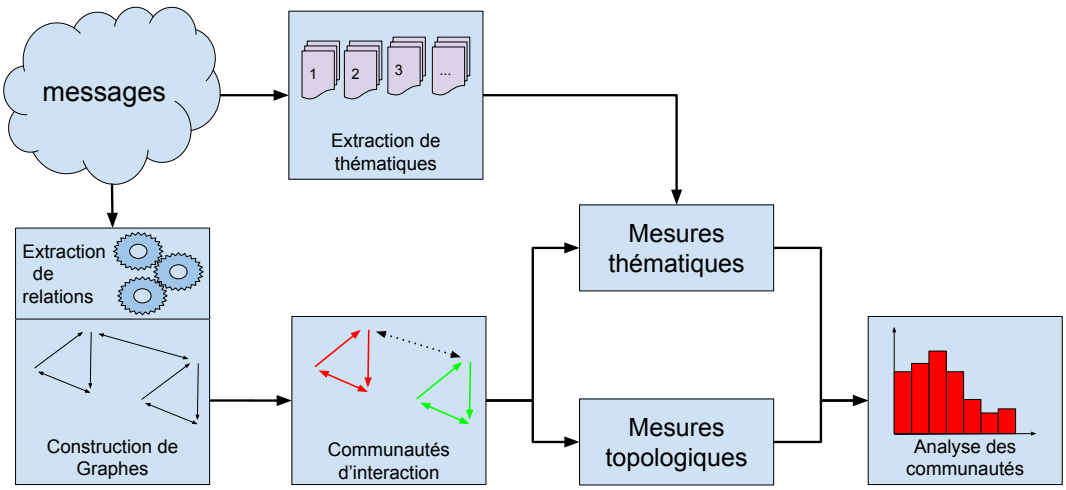



FIGURE 7.1 – Description du traitement de découverte et caractérisation de groupes d'utilisateurs

Notre contribution principale porte sur l'élaboration de mesures prenant en compte les thématiques pour développer une analyse des communautés obtenues, en exploitant les textes que les utilisateurs ont écrits et lus. Dans la littérature, cette analyse est trop souvent limitée à l'aspect topologique, c'est-à-dire à l'étude du graphe. Ces nouvelles mesures permettent d'explorer un réseau social

numérique par les communautés qu'il contient, évaluant leur cohésion et leurs influences thématiques. Elles ne remplacent pas les mesures topologiques, mais les enrichissent en amenant un second volet d'étude, complétant l'analyse des communautés.

## 7.2 Modélisation du réseau social

En analyse de réseaux sociaux, la donnée brute est constituée principalement de messages, dont certains sont porteurs d'interaction entre des utilisateurs. Nous pensons qu'il est pertinent de représenter la variété de relations entre les utilisateurs par des graphes. Dans cette section, nous introduisons notre modèle de données composé de quatre types de graphe. Nous exposons ici comment nous transformons les messages en graphes, puis comment nous détectons des groupes à partir de ces graphes.

La modélisation proposée est générique et convient à tout réseau social numérique (RSN). Afin de clarifier et concrétiser le modèle, de nombreuses références sont faites à Twitter, qui est la plus grosse plate-forme de micro-blogging, et qui a introduit des mécanismes repris par ailleurs (le mot-dièse ou *hashtag* ; la rediffusion à l'identique ou *retweet*).

### 7.2.1 Éléments informatifs de source

Sur tout réseau social numérique, un utilisateur ne peut accomplir qu'un ensemble fini d'actions. Nous introduisons ici un modèle qui se focalise sur T, l'ensemble des messages $t$ émis dans un corpus. Sur Twitter par exemple, les autres actions possibles comprennent les échanges de messages privés, l'ajout/le retrait d'abonnements (la relation de *follow*), l'ajout/le retrait à une liste, et le marquage de messages en favori (« *likes* »).

Plus largement, les messages reçus lors d'une collecte sont représentés ainsi de manière abstraite :

- **Auteur**, c'est-à-dire l'identifiant unique et le pseudonyme du compte ;
- **Corps du message**
  - Contenu. Pour les tweets, il s'agit du texte ;
  - Objets, c'est-à-dire les hashtags, liens, images ou autres items remarquables ;
  - Destinataires explicites. Il s'agit de zéro, un ou plusieurs autres comptes du même réseau social ;
- **Type** de message, par exemple tweet, message direct ou commentaire. En effet, un message peut prendre plusieurs formes différentes selon les fonctionnalités implémentées sur le réseau social.

Dans le cas de Twitter, nous nous limitons aux tweets eux-mêmes comme donnée d'entrée. Le champ *Type* est alors restreint aux trois formes d'interaction explicite : retweet, réponse ou mention. Nous utilisons la notation suivante :

- **rt** est un retweet. RT $\subset$ T est l'ensemble de tous les retweets. $rt_{i,j}^{k}$ signifie que l'utilisateur $i$ a retweeté $j$ dans le message $k$.
- **re** est une réponse. RE $\subset$ T est l'ensemble de toutes les réponses. $re_{i,j}^{k}$ signifie que l'utilisateur $i$ a répondu à $j$ dans le message $k$
- **me** est une mention. ME $\subset$ T est l'ensemble de toutes les mentions. $me_{i,j}^{k}$ signifie que l'utilisateur $i$ a mentionné $j$ dans le message $k$

Ces informations alimentent ensuite un modèle afin de faire ressortir clairement les relations entre comptes utilisateurs.

### 7.2.2 Modèles de représentation d'un réseau social

Nous représentons un réseau social tel que Twitter par différents niveaux de graphes, composés de liens de natures différentes entre les comptes utilisateurs. Selon la manière de construire les graphes, c'est-à-dire les raisons d'établir un arc entre deux nœuds-utilisateurs, plusieurs « graphes sociaux » sont construits. Dans notre modèle, nous basons notre analyse sur quatre graphes, expliqués en détail dans les paragraphes suivants :

— Graphe des amitiés, $G_F$

— Graphe d'interaction, $G_I$

— Graphe d'interaction thématiques, $G_I^\theta$

— Graphe des partages, $G_\Omega$.

Tout naturellement, le premier graphe « social » apparu dans la littérature repose sur les liens d'amitié, introduite en définition 7.2.1, caractéristique de proue du Web 2.0. Nous le notons $G_F = (U, E_F)$ où U est l'ensemble des utilisateurs d'un réseau social numérique, ou une partie de cet ensemble ; les arcs de « follow » $e \in E_F$ sont dirigés du suiveur vers le compte suivi. Sur certains réseaux où l'amitié / l'abonnement est réciproque, à chaque arc (i,j) correspond la relation réciproque (j,i).

**Definition 7.2.1.** *Amitié :* sur les réseaux sociaux numériques, « l'amitié » correspond à la fonctionnalité d'abonnement, permettant à un utilisateur de voir les publications d'un compte, sur sa page d'accueil. Il s'agit d'une relation uni-directionnelle, où le compte A est abonné au compte B.

Un second graphe, réputé mieux illustrer la réalité sociale [Lim and Datta, 2016], est basé sur les interactions, réelles et ponctuelles, décrites en définition 7.2.2. Il s'agit de $G_I = (U, E_I)$ où $E_I$ est un ensemble d'arcs représentant la quantité d'interactions entre les utilisateurs. Un arc $e_{i,j} \in E_I$ représente l'ensemble des actions émises par $i$ à destination de $j$ (par exemple, toutes les mentions), et porte un attribut de poids représentant le nombre total d'actions de $i$ vers $j$.

**Definition 7.2.2.** *Interaction :* ce terme recouvre toute action effectuée par un premier compte, touchant un second compte, à un instant précis. Par exemple, l'action de s'abonner, se désabonner, répondre, mentionner, commenter une photo, une publication, etc... d'un utilisateur. Une interaction est susceptible de déclencher une notification chez la cible de l'action.

Le texte du tweet est propice à une analyse automatique, afin de le rattacher à une thématique principale « résumant » le message. Il ne s'agit pas d'extraire le sens profond de chacun des textes, mais de quantifier les thématiques fréquentes sur des grands corpus (contenant des milliers voire des millions d'items) constitué de petits documents (les tweets sont courts, héritage de la contrainte des 140 caractères). Plusieurs algorithmes provenant du domaine du traitement automatique de la langue sont propices pour réaliser cette tâche. Sans perte de généralité, nous représentons les thématiques par la lettre $\theta$, dans un ensemble de thématiques possibles, $\Theta$. La méthode choisie attribue à chaque tweet un label de thématique $\theta$.

Les interactions visibles sur le réseau social sont portées par des messages, donc associées à du texte ; en conséquence nous pouvons créer un graphe des interactions *thématiques*. Nous notons un tel graphe $G_I^\theta = (U, E_I^\theta)$, représentant toutes les interactions concernant un thème donné. Nous obtenons donc un tel graphe $\forall \theta \in \Theta$. Ces graphes donnent la capacité d'isoler les relations concernant une thématique d'intérêt, ou bien de combiner les affinités entre nœuds parmi tout ou partie du spectre thématique possible.

Finalement, une dernière approche part d'un comportement fréquent sur les réseaux sociaux numériques : le partage. Formalisé en définition 7.2.3, il s'agit plus d'une « co-action » que d'une interaction : les utilisateurs partagent des contenus, tels que des hashtags, des liens vers des pages web,

ou des images. Nous regroupons tous ces items sous le nom « d'objets sociaux », c'est-à-dire des contenus voués au partage, à la diffusion en tant qu'objets. L'équation 7.1 définit $\Omega$ l'ensemble de ces objets.

**Definition 7.2.3.** *Partage :* deux comptes utilisateurs *partagent* un objet s'ils émettent tous deux un message contenant le même objet.

$$\Omega = [media, \text{URL}, hashtag] \tag{7.1}$$

Notre intuition est la suivante : les centres d'intérêts communs à deux utilisateurs émergent par leur représentation dans un graphe contenant à la fois les objets sociaux et les comptes utilisateurs. Certes, un graphe similaire pourrait aussi être construit seulement pour un type d'objet social ; cependant nous pensons que l'association de tous les types d'objets sociaux favorise l'apparition de relations plus fortes. Ainsi, chacun des types d'objets sociaux contribue à la construction d'un graphe bipartite de l'émission, noté $G_\Omega = (U, \Omega, E)$, reliant chaque auteur $u \in U$ aux objets sociaux $\omega \in \Omega$ qu'il émet.

TABLEAU 7.1 – Quatre graphes pour représenter un même réseau social

| Aspect | Graphe | Détails |
|---|---|---|
| Abonnement | $G_F$ | lien d'amitié, d'abonnement |
| Interaction | $G_I$ | mention, réponse ou citation |
| Interaction thématique | $G_i^\theta$ | idem, filtré par thématique |
| Partage | $G_\Omega$ | émission conjointe d'un hashtag, photo, autre |

Les quatre types de graphes sont repris en table 7.1. Ils représentent un large spectre de relations entre les comptes. L'*amitié* ou abonnement ne signifie pas forcément l'*interaction* ; de même la notion de partage permet de relever une activité simultanée entre deux comptes, que rien n'oblige à être abonnés l'un à l'autre.

Ces quatre graphes présentent l'atout d'être construits autour du même ensemble de nœuds (les comptes utilisateurs), favorisant l'application de traitements similaires, tels qu'une détection de communautés.

### 7.2.3 Détection de groupes : partitions et couvertures

En analyse de graphes, les communautés sont des ensembles de nœuds présentant des connexions plus nombreuses entre membres de la communauté, que vers d'autres communautés. Dans la littérature, plusieurs algorithmes ont été proposés pour obtenir $\Gamma$, l'ensemble des communautés $g$ détectées sur un graphe G. Comme mentionné en chapitre 4, nous distinguons les partitions, qui attribuent à chaque nœud une unique communauté, et les couvertures, qui autorisent le recouvrement entre communautés : un nœud peut simultanément faire partie de plusieurs groupes.

Pour calculer la partition d'un graphe, nous retenons Louvain [Blondel et al., 2008], qui est rapide même sur des graphes grands (plusieurs centaines de milliers de nœuds), et prend en compte la pondération des arcs. C'est une méthode stable : malgré une initialisation aléatoire, les résultats finaux sont similaires.

La complexité et le manque de vérité-terrain pour la validation sont deux grands freins pour les algorithmes de couverture, autorisant la multi-affiliation. Ces méthodes donneraient plus de sens sur les « ego-networks » : les réseaux centrés sur un unique nœud d'appartenance multiple. Il ne s'agit alors pas d'identifier les groupes d'utilisateurs qui structurent le réseau, mais d'étudier un compte par son entourage.

Qu'il s'agisse d'une partition ou d'une couverture, les communautés obtenues doivent ensuite être caractérisées : il faut mesurer si elles sont réellement des communautés, ou seulement des ensembles de nœuds faiblement reliés entre eux.

### 7.2.4 Mesures topologiques de qualité des communautés

L'évaluation de la force de liaison interne des communautés est un problème déjà bien documenté dans la littérature. Quelques mesures ont déjà été introduites en chapitre 4. Nous les rappelons brièvement ici :

— **Densité interne**, qui évalue la quantité de liens internes ;

— **TPR** pour *Triad Proportion Ratio*, la proportion de triangles au sein des communautés ;

— **Conductance**, correspondant à la proportion des liens émis par les membres d'un groupe, dirigés vers l'extérieur ;

— **Modularité**, qui met en relation les densités internes des groupes par rapport à un graphe aléatoire uniforme ; elle mesure la qualité de la partition d'un graphe dans sa globalité.

Ces mesures caractérisent des aspects très différents d'un groupe : la densité interne évalue la quantité de liens internes, c'est-à-dire entre les membres d'une même communauté. La *TPR* calcule la proportion de triangles au sein des communautés, ce qui mesure à quel point les membres y sont intégrés (la présence « d'amis » qui se connaissent les uns les autres est un facteur de force du groupe). La conductance permet de considérer à quel point un groupe est lié à son environnement (avec des valeurs allant de 0, déconnecté, à 1, exclusivement connecté à l'extérieur sans aucun lien interne). La modularité ne donne pas de valeur aux communautés, mais à la partition ; il n'est donc pas possible de caractériser une seule communauté par sa modularité.

Grâce à ces mesures, la quantité d'interaction au sein d'un groupe est correctement évaluée ; pour ajouter par la suite la dimension sémantique, il faut tout d'abord analyser le texte des messages échangés.

### 7.2.5 Extraction de la thématique des textes

La caractérisation thématique des communautés nécessite une étape préalable d'analyse des textes, afin de détecter les thématiques présentes dans les échanges entre comptes utilisateurs. Dans ces paragraphes, nous exposons quelques méthodes et expliquons notre choix pour l'une d'entre elles.

Nous définissons $\Theta$ l'ensemble des N thématiques $\theta_1, \ldots, \theta_N$, incluant un thème « Aucun » pour représenter les documents inclassables. Il existe plusieurs implémentations de la détection de thématiques dans la littérature, adaptées à des applications spécifiques ; ces implémentations sont présentées par une opposition entre classification supervisée et non-supervisée. La première approche nécessite des thématiques pré-établies, telles que « Politique / Culture / Sport » issues de l'analyse d'articles de presse par exemple. La seconde approche, non-supervisée, regroupe les textes selon une distance ou similarité ; une thématique correspond alors à un groupe de documents. Vu le spectre des applications de l'analyse des réseaux sociaux, un *a priori* sur les thématiques présentes nous semble difficile à obtenir ; nous y préférons une méthode adaptable et non-supervisée, par exemple LDA (*Latent Dirichlet Allocation*) [Blei et al., 2003] .

Dans la méthode LDA, une thématique $\theta$ est un vecteur, similaire et représentatif des textes analysés : un vecteur $tf.idf$, qui représente la fréquence d'un mot dans un document ($tf$) pondérée par sa diffusion dans l'ensemble du corpus ($idf$). Dans cet espace de projection choisi, une similarité sémantique est définie entre un texte et une thématique ; un choix usuel consiste en la similarité $cosinus$ entre vecteurs $tf.idf$, qui résulte en un score compris entre 0 (dissimilarité totale) et 1 (identité). Ainsi, les thématiques sont des groupes de documents très similaires entre eux.

Les vecteurs $tf.idf$ étant de grande dimension (nombre de mots dans le vocabulaire), chaque texte $t$, par sa position dans l'espace de projection, est approximé par une somme pondérée de

quelques thématiques différentes : à chaque thématique $\theta_i$ est associé un poids, $w_i$, calculé comme la similarité sémantique entre le texte $t$ et le thème $\theta_i$. Une valeur élevée de $w_i$ signifie que le texte $t$ est similaire à la thématique $\theta_i$.

$$\theta_t = \theta_i \in \Theta \text{ tel que } \forall \theta_j \in \Theta, w_i^t \geq w_j^t \tag{7.2}$$

Parmi ces thématiques candidates, seule la thématique principale est retenue comme étiquette pour le texte $t$ : celle de similarité maximale, comme défini en équation 7.2. Il s'agit d'une hypothèse assez forte, à mettre en relation avec les éléments suivants :

— la quantité de messages analysés. En effet, la conservation des vecteurs $tf.idf$, ou même des N poids en relation aux N thématiques représente un grand volume de données, plus difficile à exploiter ;

— l'objectif final consiste à déterminer les thématiques favorites pour chaque utilisateur, puis pour chaque groupe. Ce double seuillage éliminerait, de toute façon, les thématiques secondaires.

À ce stade, chaque message est étiqueté par une thématique ; l'agrégation des thématiques sur plusieurs messages, provenant d'un même utilisateur voire d'un même groupe d'utilisateurs, nécessite une étape de définition des notations utilisées.

$T_g$, défini en équation 7.3, est l'ensemble de tous les messages émis par les membres d'un groupe $g$. L'auteur d'un message $t$ est le compte utilisateur $u(t)$. Un utilisateur mentionné dans un message en est le destinataire, noté $dest(t)$ ; un message n'a pas forcément de destinataire explicite. Pour qu'un message soit positionné dans un groupe, le message doit être émis par un membre du groupe. Bien évidemment, $u(t)$ et $dest(t)$ sont des éléments de V l'ensemble des nœuds du graphe G.

$$T_g = \{t, u(t) \in g\} \tag{7.3}$$

De manière similaire, $T_\theta$ est défini comme l'ensemble de tous les tweets de thématique $\theta$. Chaque utilisateur est étiqueté par la thématique majoritaire de ses messages. Pour ce faire, l'ensemble des messages $T_u$ écrits par le compte $u$ est analysé. Pour conserver une représentation simple, nous ne retenons que la thématique principale pour chaque compte, notée $\theta_u$ dans l'équation 7.4. De la même façon, une thématique $\theta_g$ est attribuée à chaque groupe d'utilisateurs $g$, suivant l'équation 7.5.

$$\theta_u = freqmax_{\theta_i} T_u \tag{7.4}$$

$$\theta_g = freqmax_{\theta_i} \{\theta_u \, \forall u \in g\} \tag{7.5}$$

Finalement, nous notons $U_\theta$ l'ensemble des utilisateurs de thématique principale $\theta$.

Dans cette section, nous avons proposé la modélisation d'un réseau social au travers de quatre types de graphes ; dans un second temps nous avons rappelé les techniques issues de l'état de l'art concernant la détection de communautés dans un graphe, les mesures topologiques de qualité de telles communautés, et les techniques d'extraction de thématiques des textes. Tous les éléments sont rassemblés pour poursuivre l'analyse des communautés détectées, au travers du prisme des thématiques abordées.

## 7.3 Mesures de la cohésion sémantique d'un groupe

Nous souhaitons caractériser les communautés d'utilisateurs de réseaux sociaux ; les mesures topologiques issues de l'analyse de graphe donnent des éléments pour calculer la force d'interaction interne d'un groupe, mais ne prennent pas en compte le contenu des textes émis et échangés, à la base de la relation d'interaction entre les comptes. Pour pallier ce manque, nous introduisons dans cette section un ensemble de mesures pour quantifier la cohésion sémantique des groupes.

### 7.3.1 Scores de relation entre thématiques et groupes

Inspirées de la précision et du rappel, deux mesures thématiques $\xi$ et $\rho$ sont définies en équations 7.6 à 7.11, qui mesurent à quel point les groupes correspondent à la répartition des thématiques. L'intuition est de mesurer, d'une part combien de membres d'un groupe sont intéressés par une même thématique, et d'autre part, à quel point un groupe rassemble les utilisateurs intéressés par un thème.

#### 7.3.1.1 Expertise $\xi$ d'un groupe sur une thématique

L'*Expertise* $\xi$, inspirée de la précision, donne lieu à deux variantes. La première, notée $\xi_u$ et définie en équation 7.7, représente la quantité d'utilisateurs dans un groupe qui sont « experts » sur une thématique donnée : il s'agit de la proportion des membres du groupe, dont la thématique principale est la même. Pour continuer le parallèle avec la précision dans le domaine de l'apprentissage automatique, le classifieur réalise la détection des groupes, qui doivent correspondre à une vérité-terrain, ici la thématique des textes émis par les membres des groupes.

La deuxième variante, notée $\xi_t$ et définie en équation 7.8, utilise la quantité de messages d'une même thématique émis par les membres du groupe. Plutôt que de dénombrer les utilisateurs, il s'agit de quantifier les contenus.

$$\xi \,:\, \Theta,\, \Gamma \rightarrow [0,1] \subset \mathbb{R} \tag{7.6}$$

$$\xi_u(\theta, g) = \frac{\#\{u \in g, \theta_u = \theta\}}{\#g} \tag{7.7}$$

$$\xi_t(\theta, g) = \frac{\#\{t \in \mathrm{T}_g, \theta_u = \theta\}}{\#\mathrm{T}_g} \tag{7.8}$$

Dans les deux cas, $\xi(g) = 1$ caractérise un groupe parfaitement cohésif ; des valeurs proches de zéro seront liées à un groupe très dispersé sémantiquement. Cette mesure est favorable aux groupes-thématiques, c'est-à-dire des groupes dont les membres n'abordent qu'un seul thème.

#### 7.3.1.2 Représentativité $\rho$ d'un groupe sur une thématique

Construite en miroir, la *Représentativité* $\rho$, inspirée du rappel et définie en équations 7.9 à 7.11, est la proportion d'utilisateurs dans la globalité du réseau, qui sont intéressés dans la même thématique et présents dans le groupe. Cette mesure est focalisée sur la distribution d'une thématique entre les groupes : est-ce qu'un groupe est le centre du thème, ou seulement un contributeur secondaire ? Des valeurs proches de zéro suggèrent que beaucoup de groupes partagent le même intérêt pour cette thématique.

Ici aussi, deux supports sont disponibles pour calculer la mesure : le nombre d'utilisateurs actifs, résultant en $\rho_u$, en équation 7.10, ou la quantité de messages émis, résultant en $\rho_t$, en équation 7.11. Le premier repose sur un centrage sur l'individu : un groupe n'est cohésif que si tous prennent part au débat ; le second privilégie l'activité : le mutisme des uns est compensé par la loquacité des autres, au risque de distinguer les comptes trop bavards.

$$\rho \,:\, \Theta,\, \Gamma \rightarrow [0,1] \subset \mathbb{R} \tag{7.9}$$

$$\rho_u(\theta, g) = \frac{\#\{u \in g, \theta_u = \theta\}}{\#\mathrm{U}_\theta} \tag{7.10}$$

$$\rho_t(\theta, g) = \frac{\#\{t \in \mathrm{T}_g, \theta_t = \theta\}}{\#\mathrm{T}_\theta} \tag{7.11}$$

#### 7.3.1.3 Cohésion d'un groupe

À partir des formules des équations 7.12 et 7.13, nous allégeons la notation en $\xi(g)$ et $\rho(g)$ pour évaluer la présence de la thématique principale $\theta_g$ dans le groupe $g$ ; cette présence est mesurée par le nombre d'utilisateurs. Bien qu'il soit possible de calculer ces valeurs pour chaque thématique, la cohésion du groupe est forcément maximale sur la thématique principale. De plus, nous préférons mettre l'accent sur la participation de chaque membre du groupe, privilégiant $\xi_u$ et $\rho_u$. Ainsi chaque groupe est décrit par seulement deux valeurs de cohésion sémantique, plutôt qu'une par thématique potentielle.

$$\xi(g) = \xi_u(\theta_g, g) \tag{7.12}$$

$$\rho(g) = \rho_u(\theta_g, g) \tag{7.13}$$

Ces mesures sont faciles à manipuler, puisqu'elles produisent des scores assimilables à des pourcentages : de zéro, faible, à 1 = 100%, excellent. En revanche, elles considèrent les thématiques comme des étiquettes atomiques, totalement différentes les unes des autres : or, deux thématiques proches peuvent co-exister au sein d'un groupe.

Le fait de baser les calculs de ces mesures soit sur le nombre de textes, soit sur le nombre d'utilisateurs, donne lieu à des expériences ultérieures. Il ne s'agit pas d'un choix anodin : le nombre de messages dont l'émission est automatique est élevé sur Twitter, et résulte en un déséquilibre en faveur de ces robots. Lorsque le cas d'application requiert l'identification de groupes de robots, $\xi_t$ est adéquat ; hors de cette situation, nous pensons préférable d'utiliser $\xi_u$ comme score de cohésion thématique.

### 7.3.2 Cohésion considérant une similarité entre thématiques

Nous proposons un autre indicateur de cohésion thématique, nommé $\xi_{sim}$. Il considère une similarité entre les thématiques, prenant mieux en compte la variabilité des contenus des $\theta$ : deux thèmes sont différents si les centres de clusters de documents sont éloignés l'un de l'autre. Ainsi deux thématiques (très) proches ne devraient pas être considérées comme totalement différentes : un groupe reste cohérent s'il aborde deux thèmes proches. En conséquence nous proposons une adaptation de la mesure de cohésion, introduite en équation 7.14.

$$\xi_{sim}(g) = \frac{1}{\#g} \sum_{u \in g} (sim(\theta_u, \theta_g))^2 \tag{7.14}$$

Dans cette équation, $sim$ est une mesure de similarité entre thématiques, allant de 0 (différence totale) à 1 (même thème). Des exemples de telles mesures de similarité incluent la « *cosine similarity* » (valeur absolue du cosinus entre deux vecteurs), ou d'autres basées sur des distances (euclidienne par exemple). Dans notre cas, avec des vecteurs $tf.idf$, le choix usuel est la similarité *cosinus*. La passage au carré de la valeur de similarité en équation 7.14 diminue le poids des thèmes « faiblement différents » dans le calcul.

Les valeurs de $\xi_{sim}$ obtenues sont aussi inclues dans l'intervalle $[0, 1]$, mais réparties différemment de celles de $\xi$ : à l'importance du thème principal sont ajoutées les similarités et importances des thèmes secondaires. Il en résulte des scores globalement plus élevés, qui ne sont pas directement comparables avec les $\xi$ précédemment introduits.

### 7.3.3 Score de pertinence thématique d'un groupe

Intuitivement, les occurrences des thématiques se comportent comme des termes dans un document : quelques thèmes sont fréquents, présents partout, alors que d'autres sont beaucoup plus spécifiques, mentionnés dans un seul groupe. Suivant l'idée derrière la représentation en $tf.idf$, nous

proposons $\theta f.igf$ pour « *topic frequency, inverse group frequency* », dans les équations 7.15 et 7.16. Ce score caractérise la présence d'une thématique $\theta$ dans un groupe $g$. Pour rappel dans la notation, $\Gamma$ est l'ensemble des communautés $g$ détectées.

Le premier terme $\theta f$ mesure la fréquence d'apparition d'une thématique dans un groupe, et est égal à $\xi$. Le second terme $igf$ relativise la fréquence d'un thème en comptant le nombre de groupes dans lequel il est majoritaire.

$$\theta f = \frac{\#\{u \in g, \theta_u = \theta\}}{\#g} \tag{7.15}$$

$$igf = log \frac{\#\Gamma}{1 + \#\{g \in \Gamma, \theta_g = \theta\}} \tag{7.16}$$

Le score final est le produit de $\theta f$ et $igf$. Afin d'éviter les divisions par zéro, nous ajoutons 1 au dénominateur, à $igf$. Des valeurs élevées, supérieures à 1, signifient qu'une thématique $\theta$ est très fréquente dans un groupe $g$, et plutôt rare parmi les autres groupes d'utilisateurs. Des valeurs basses, proches de zéro, signifient soit que le thème est très fréquent, globalement, parmi la population ; soit que cette thématique principale est malgré tout très rare dans le groupe considéré.

### 7.3.4 Synthèse des mesures introduites

La table 7.2 dresse un résumé des mesures introduites et rappelle le domaine des scores possibles ; elle propose aussi des valeurs qui aident à qualifier les scores obtenus pour des groupes détectés dans un graphe social, à leur associer une signification.

TABLEAU 7.2 – Description des indicateurs

| | $\xi$ | $\rho$ | $\xi_{sim}$ | $\theta f.igf$ |
|---|---|---|---|---|
| Nom | Expertise | Représentativité | Expertise (*bis*) | Pertinence |
| Domaine | $[0,1]$ | $[0,1]$ | $[0,1]$ | $[0,+\infty[$ |
| Valeurs élevées | >0.9 | $>10^{-2}$ | >0.86 | >0.8 |
| Valeurs faibles | <0.2 | $<10^{-6}$ | <0.75 | <0.5 |

Les valeurs de $\xi$ seront souvent plus faibles pour de grands groupes, la probabilité se réduisant que tous les membres soient intéressés par exactement le même thème. Les distributions des valeurs de $\theta f.igf$ et $\rho$ dépendent directement du nombre de groupes et de comptes intéressés par un thème donné. Ainsi, un corpus plus grand résulte en des scores $\theta f.igf$ globalement plus bas pour tous les groupes. Cela n'impacte pas la comparaison des cohésions de groupes au sein d'un même réseau.

Nous proposons d'illustrer ces mesures dès la section suivante, sur un jeu de données fourni en exemple, *ArtsTweets*.

## 7.4 Illustration par un exemple artificiel : *ArtsTweets*

Nous proposons un exemple fictif, de taille modeste, pour illustrer l'intérêt de la contribution. Le corpus, constitué de 34 messages émis par 19 comptes, est placé en Annexe B.2, dans la table B.1. Chacun s'y exprime à propos de musique, de cinéma, de théâtre ; un compte fait la promotion d'un site web ; quelques-uns souhaitent aller (ensemble ?) à un festival.

Sur ces tweets, un processus de reconnaissance d'interactions, de thématique et de partage est appliqué, alimentant la table B.2 en Annexe B.2. La colonne des *Interactions* sert de donnée d'entrée pour construire le graphe $G_I$, reliant les auteurs de tweets à leurs destinataires explicitement mentionnés dans le texte ; le graphe ainsi obtenu est illustré en figure 7.2. La thématique la plus fréquente dans les émissions d'un compte résulte en une couleur spécifique pour le nœud ; les thèmes reconnus

humainement (*ciné, théâtre, jazz, métal* sont remplacés par les étiquettes $\theta_1, ... \theta_4$. Dans les figures 7.2 à 7.4, nous conservons la même spatialisation : elle ne porte pas de sens en soi, il s'agit uniquement d'une aide à la lecture du graphe.

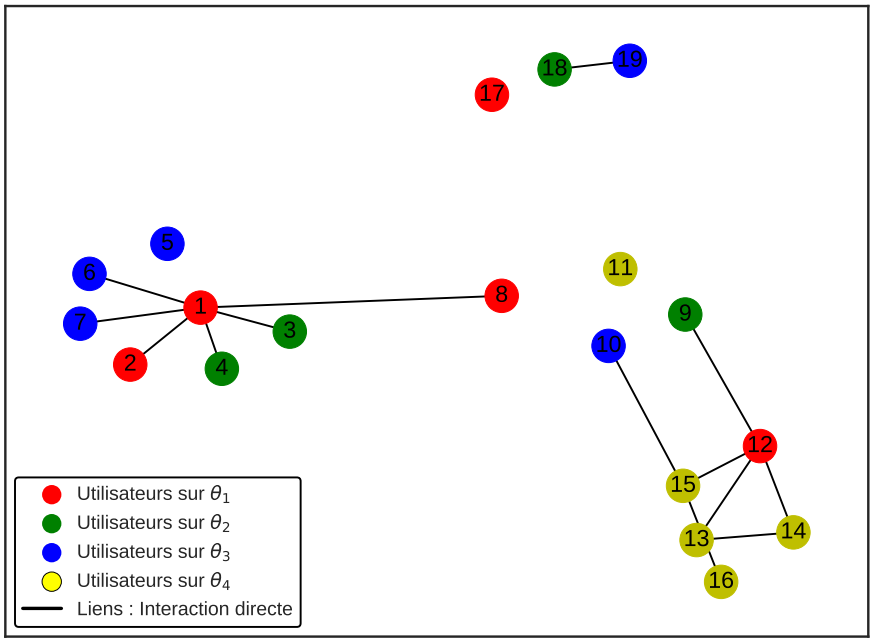



FIGURE 7.2 – Visualisation des interactions dans *ArtsTweets*

De manière similaire, la figure 7.3 présente les relations entre les comptes et les objets sociaux (nœuds en gris) qu'ils ont émis, vue comme un graphe bipartite. La projection de ce graphe sur l'ensemble des utilisateurs place une relation directe entre ceux qui ont émis les mêmes contenus.

Finalement, la figure 7.4 combine les deux graphes précédents, en conservant la même spatialisation. Cette combinaison s'obtient en deux étapes : tout d'abord, la projection du graphe bipartite sur les nœuds utilisateurs retire les nœuds de contenus ; le graphe ainsi obtenu est uni (union des listes d'arcs) avec le graphe des interactions. Ce graphe final est composé de deux parties connexes. L'attribution des nœuds à des communautés, obtenue par optimisation de la modularité via l'algorithme Louvain [Blondel et al., 2008], est proposée visuellement par des ellipses et une étiquette de groupe.

Afin de couvrir l'ensemble des mesures introduites, il est nécessaire de disposer des similarités entre les thématiques mentionnées par les comptes. Ainsi, la table 7.3 propose des valeurs artificielles, car la méthode *LDA* ne fonctionne pas bien avec si peu de documents en entrée ; pour simplifier la lecture, puisque la matrice est symétrique, nous la représentons en matrice triangulaire. Chaque thème est exactement similaire à lui même ; la plupart des thèmes sont éloignés les uns des autres (similarité de 0.1), à l'exception de $\theta_2$ et $\theta_3$, qui sont proches (similarité de 0.9).

Pour calculer les scores $\xi_u$, $\rho_u$, $\xi_t$, $\rho_t$, $\xi_{sim}$ et $\theta f.igf$, nous dénombrons le nombre de comptes intéressés par chacune des thématiques, pour chaque groupe. Ces valeurs sont résumées en table 7.4. Par exemple, il y a 4 membres du *Groupe_3* qui sont intéressés par la thématique $\theta_4$.

La thématique majoritaire de chaque groupe est calculée, marquée en gras en table 7.4 et inscrite dans la colonne $\theta_g$ en table 7.6. Lorsque plusieurs thèmes arrivent à égalité au sein d'un groupe, par convention, celui d'indice minimal est retenu : par exemple, $\theta_1$ pour le *Groupe_2*.

Les mesures thématiques sont ensuite calculées selon leurs définitions, et sont visibles en table

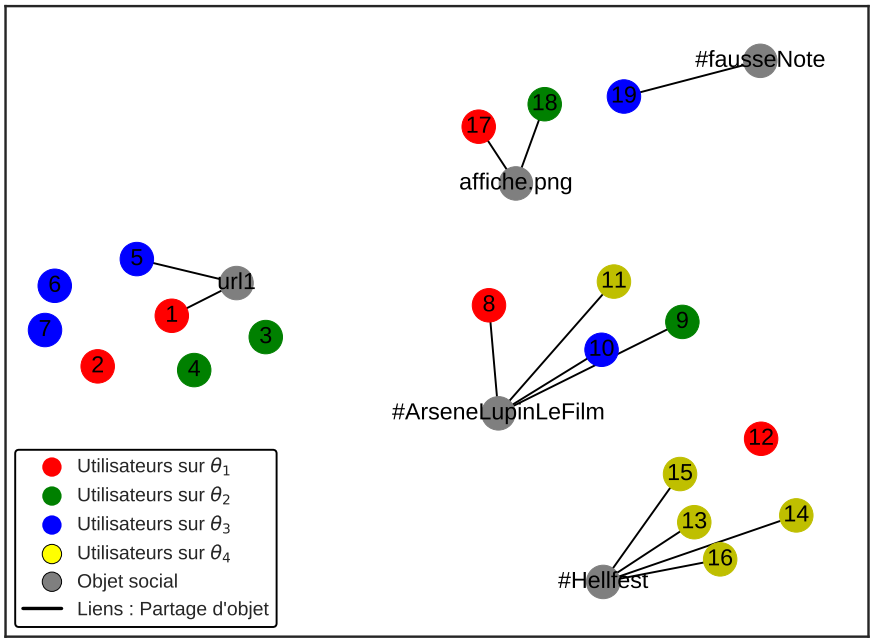



FIGURE 7.3 – Visualisation des partages d'objets sociaux dans *ArtsTweets*

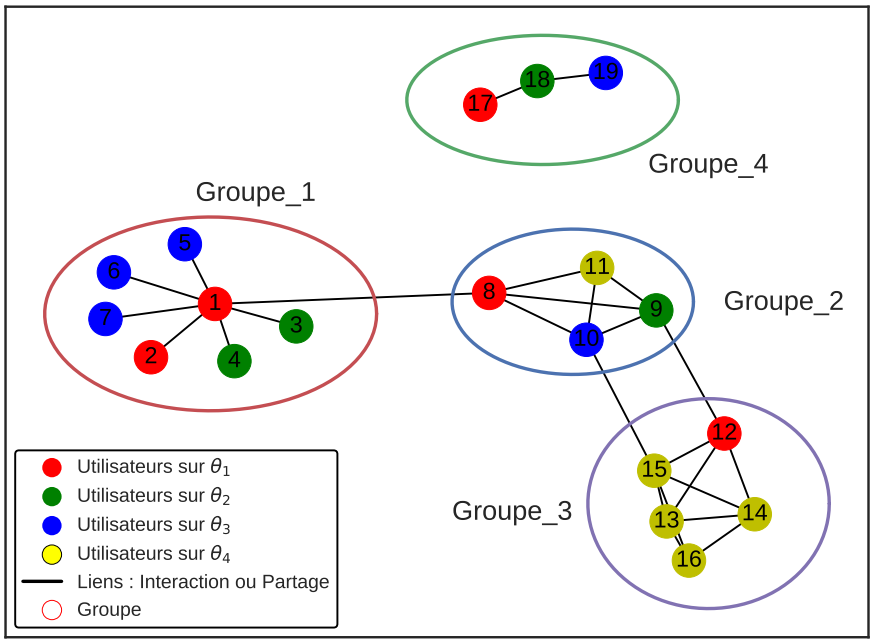



FIGURE 7.4 – Visualisation des communautés détectées

TABLEAU 7.3 – Similarités entre les thématiques

| Similarité | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|---|
| $\theta_1$ | 1 | | | |
| $\theta_2$ | 0.1 | 1 | | |
| $\theta_3$ | 0.1 | 0.9 | 1 | |
| $\theta_4$ | 0.1 | 0.1 | 0.1 | 1 |

TABLEAU 7.4 – Répartition des centres d'intérêts par groupe, en nombre d'utilisateurs

| | Groupe_1 | Groupe_2 | Groupe_3 | Groupe_4 | Total |
|---|---|---|---|---|---|
| $\theta_1$ | 2 | **1** | 1 | **1** | 5 |
| $\theta_2$ | 2 | 1 | 0 | 1 | 4 |
| $\theta_3$ | **3** | 1 | 0 | 1 | 5 |
| $\theta_4$ | 0 | 1 | **4** | 0 | 5 |
| Total | 7 | 4 | 5 | 3 | 19 nœuds |

7.6. Pour le Groupe_1, $\#g = 7$. La table 7.4 permet de lire les données nécessaires : par exemple, il y a 3 membres intéressés par le même thème $\#\{u, u \in Groupe\_1, \theta_u = \theta_3\} = 3$ ; ce qui permet de calculer $\xi_u(Groupe_1) = 0.43$. De plus, le nombre total de comptes actifs sur cette thématique est $U_{\theta_3} = 5$, d'où nous calculons $\rho_u(Groupe\_1) = 0.60$. Ensuite, l'indicateur de pertinence est décomposée en deux parties : d'une part, $\theta f = \frac{\#\{u, u \in Groupe\_1, \theta_u = \theta_3\}}{7} = 0.43$. Le second terme vaut $igf = \log \frac{\#\Gamma}{1+\#\{g, \theta_g = \theta\}} = 0.30$, d'où nous obtenons $\theta f.igf = 0.13$. Finalement, la valeur de $\xi_{sim}$ est calculée en prenant en compte chacun des membres du groupe, ainsi que la proximité entre les thématiques. Pour les 7 membres du *Groupe_1*, une valeur de 3 provient du thème majoritaire $\theta_2$ ; $2 \times 0.9 = 1.8$ proviennent des membres verts ; $2 \times 0.1 = 0.2$ proviennent des membres rouges : un score final de $\xi_{sim} = \frac{3+1.8+0.2}{7} = 0.71$ est obtenu.

La même démarche est appliquée pour calculer $\xi_t$ et $\rho_t$, mesures basées sur le nombre de messages émis par le groupe. Pour accompagner le calcul, la table 7.5 fournit les informations.

TABLEAU 7.5 – Répartition des centres d'intérêts par groupe, en nombre de tweets

| | Groupe_1 | Groupe_2 | Groupe_3 | Groupe_4 | Total |
|---|---|---|---|---|---|
| $\theta_1$ | 8 | 4 | 1 | 2 | 15 |
| $\theta_2$ | 2 | 2 | 0 | 2 | 6 |
| $\theta_3$ | 3 | 1 | 0 | 1 | 5 |
| $\theta_4$ | 0 | 2 | 6 | 0 | 8 |
| Total | 13 | 9 | 7 | 5 | 34 messages |

La table 7.6 montre donc les scores thématiques et topologiques obtenus pour chacun des groupes, permettant l'analyse suivante. Le *Groupe_1*, de faible densité $d_{int} = 0.29$, centré autour d'un unique nœud, est dispersé sur trois thèmes, résultant en un faible $\xi_u = 0.43$. La proximité de deux thématiques corrige cette dispersion apparente, avec $\xi_{sim} = 0.73$. L'intense activité du nœud central, assimilable à du spam, lui attribue un $\xi_t = 0.62$ relativement élevé. La grande taille du groupe, en comparaison du reste de *ArtsTweets* lui confère cependant une majorité sur le thème $\theta_3$ ($\rho_u = 0.60$).

Le *Groupe_2* est une clique : ainsi les membres y sont fortement liés (TPR maximal), obtenant une densité interne de 1.0. Cependant, nous leur avons attribué des thématiques différentes, illustrant une divergence de centres d'intérêt, confirmée par le faible $\xi_{sim}$ obtenu. Il en résulte de faibles scores de cohésion $\xi_u = 0.25$ et de pertinence $\theta f.igf = 0.03$. Sa position de pont entre les *Groupe_1* et 3 lui confère la plus grande valeur de conductance de l'exemple.

Le *Groupe_3* est presque une clique, qui montre une forte cohésion thématique autour de $\theta_4$ (en jaune), atteignant ainsi un score $\xi_u = 0.80$. Comme il s'agit de l'un des deux seuls groupes à

TABLEAU 7.6 – Mesures obtenues par les groupes dans *ArtsTweets*

| Informations | | Thématiques | | | | | | Topologiques | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Groupe | $\theta_g$ | $\xi_u$ | $\rho_u$ | $\xi_t$ | $\rho_t$ | $\theta f.igf$ | $\xi_{sim}$ | $d_{int}$ | TPR | C |
| 1 | $\theta_3$ | 0.43 | 0.60 | 0.62 | 0.53 | 0.13 | 0.71 | 0.29 | 0.0 | 0.08 |
| 2 | $\theta_1$ | 0.25 | 0.25 | 0.44 | 0.27 | 0.03 | 0.33 | 1.0 | 1.0 | 0.20 |
| 3 | $\theta_4$ | 0.80 | 0.80 | 0.86 | 0.75 | 0.24 | 0.82 | 0.80 | 1.0 | 0.10 |
| 4 | $\theta_1$ | 0.33 | 0.20 | 0.40 | 0.33 [1] | 0.04 | 0.66 | 0.66 | 0.0 | 0.0 |

mentionner cette thématique $\theta_4$, le *Groupe_3* atteint aussi un $\rho_u$ élevé, et se distingue par sa pertinence relativement élevée $\theta f.igf = 0.24$, bien que la taille du corpus ne permette pas d'obtenir de score particulièrement haut.

Enfin, le *Groupe_4* est faiblement relié avec seulement deux arcs internes, pas de triangles (TPR = 0) ; il est isolé du reste, avec une conductance nulle, et n'obtient de majorité sur aucun thème, obtenant un faible $\xi_u = 0.33$, et un $\xi_t = 0.40$ comparable.

La différence de support pour les deux variantes de $\xi$ et $\rho$, qui sont basées soit sur le nombre d'utilisateurs, soit sur le nombre de textes, résulte en une tendance similaire. Une exception notable à signaler provient de la sur-activité du compte 1, augmentant mécaniquement le poids du groupe sur cette thématique. Dans *ArtsTweets*, il s'agit de l'unique comportement de spam présent ; avec des données réelles, ce type de compte est plus fréquent.

À l'issue de cette illustration, les mesures topologiques donnent une première vue du réseau : le Groupe_1 est très peu dense ; le *Groupe_4* ne contient aucun triangle. L'information est claire : ces communautés ne sont pas portées par l'interaction entre leurs membres.

Ces mêmes mesures topologiques distinguent les *Groupe_2* et *Groupe_3*, par leurs bonnes densités et TPR. Une fois reconnue la force de liaison interne, il est possible de se pencher sur la conductance : elle est parfois très élevée pour des communautés faiblement liées en interne. En l'occurrence, la conductance donne au *Groupe_2* la meilleure place comme relais de l'information dans le réseau.

En revanche, il faut inclure l'analyse sémantique pour observer la dispersion totale du *Groupe_2*, quantifiée par les faibles $\xi_u$ et $\theta f.igf$. Il en va de même pour identifier que le *Groupe_3* concentre l'écrasante majorité des *fans* de la thématique $\theta_4$.

La combinaison des deux types de mesures donne une vue complète du réseau, en tirant profit de la structure du graphe ainsi que des informations de thématique provenant de l'analyse des textes pour caractériser les communautés trouvées sur *ArtsTweets*. Ainsi le *Groupe_3* se distingue comme fortement soudé (densité et TPR élevés), particulièrement cohésif (forts $\xi_u$ et $\xi_t$), ayant un impact important sur sa thématique ($\rho_u$ et $\rho_t$ élevés) et pertinent (bon score $\theta f.igf$).

## 7.5 Discussion sur les communautés

La détection de communautés dans un graphe est, en soi, un problème difficile. L'accès à des données sociales réelles nous encourage à mieux explorer et comprendre ce que les communautés sont vraiment. En cela, les RSN présentent une opportunité pour travailler sur ce type de données. Cela déclenche aussi quelques réflexions sur les limites de ce modèle tant sur l'utilisation des graphes et de la notion de communautés, que sur les mesures que nous venons d'introduire.

### 7.5.1 Limites de la représentation en graphes

La vision d'un réseau social en recourant à l'objet *graphe* s'est naturellement imposée à nos yeux ; bien qu'elle permette d'extraire des communautés qui font sens, cette vision est porteuse de quelques

biais et limites :

— **Temporalité.** Des modèles de graphes dynamiques existent, et permettent de représenter des actions telles que la publication de messages. Il est plus compliqué de représenter la dynamique de la proximité entre deux utilisateurs ainsi que de faire évoluer les communautés au fil du temps.

— **Complétude.** Comme les données proviennent de réseaux sociaux numériques réels, nous ne disposons que d'une partie des données (quelques utilisateurs, pas tous), sur une période donnée (ni avant, ni après), sur une plate-forme donnée (par exemple, Twitter, mais pas d'autres RSN), et sur un format donné (par exemple, des tweets, pas des messages directs). En conséquence, nous ne pouvons que supposer l'intensité de l'interaction entre deux comptes : le jeu de données étudié est donc souvent incomplet.

— **Réalité des arcs.** Le modèle de graphe retenu n'est jamais une solution miracle ; il est malheureusement fréquent de l'oublier. Par exemple, la relation d'abonnement est souvent présentée comme impliquant une confiance stable et durable entre deux comptes. Cependant des travaux ont déjà critiqué ce type précis de graphe [Lim and Datta, 2016], certains abonnements n'étant pas aussi *intenses* que d'autres : ajout par proximité réelle, par réputation, à l'occasion d'une rencontre ponctuelle... Certes, par la suite, certains de ces arcs sont porteurs d'une relation régulière, mais parfois aussi d'aucune relation ni échange futur.

À partir de ces graphes, nous avons détecté des communautés par partitionnement. Cependant, dans la vie réelle, les individus appartiennent à plusieurs groupes : nous parlons de communautés avec multi-affiliation. Travail, amis, famille, clubs sont autant de communautés susceptibles de transparaître dans les réseaux sociaux numériques. Actuellement, les algorithmes de détection de *couvertures* répondant à ce problème sont difficiles à évaluer, à cause notamment de la difficulté d'obtention d'une vérité-terrain.

Ces quelques éléments font que le choix d'une représentation reste ouvert au débat ; notre proposition de construire les graphes $G_F$, $G_I$, $G_I^\theta$ et $G_\Omega$ permet de représenter différents aspects de la relation sociale numérique. Au besoin, la combinaison de ces graphes renforce les relations entre les comptes. De plus, les traitements de détection de communautés et de calcul de mesures de cohésion thématique sont réalisables sur chacun des graphes construits.

### 7.5.2 Limites des mesures de cohésion

Dans ce chapitre, nous avons également introduit des mesures thématiques, qui attribuent un poids d'importance d'une communauté sur un sujet donné (grâce à $\rho$), et évaluent la cohésion interne d'un groupe (via $\xi$), c'est-à-dire le nombre de comptes partageant un même centre d'intérêt. Ces importances sont basées sur le nombre de comptes utilisateurs ($\xi_u$ et $\rho_u$), ou sur la quantité de messages émis ($\xi_t$ et $\rho_t$) : les deux premières mesures mettent en avant la participation de tous les membres du groupe, au risque d'attendre en vain des utilisateurs-lecteurs, qui s'expriment généralement très peu ; les secondes mesures, ($\xi_t$ et $\rho_t$), récompensent la production de contenus, centrant le groupe sur ses utilisateurs les plus actifs.

Ces nouvelles mesures viennent compléter d'autres caractéristiques, plus classiques, basées sur la topologie du graphe et la taille des communautés obtenues. Ainsi, tant la force d'interaction interne, topologique, que le partage d'affinités, thématique, sont mesurés. De plus, si une similarité entre thématiques est disponible, la mesure $\xi_{sim}$ permet de l'exploiter.

Une autre approche explorée, ($\theta f.igf$), inspirée de $tf.idf$, évalue la spécificité et l'importance d'un thème dans un groupe, prenant en compte le poids de la thématique parmi l'ensemble du jeu de données. L'analyse de ce score ouvre quelques perspectives sur un corpus à la fois textuel et social, en distinguant des groupes très focalisés sur des thématiques peu courantes.

Il faut relever une sensibilité des mesures introduites à divers facteurs : les nombres de nœuds, de groupes et de thématiques dans le jeu de données ont un impact certain sur les valeurs prises par

les mesures, ce qui peut modifier de manière importante les échelles de représentation. Sur un grand jeu de données, cela résulte en des valeurs de $\rho$ globalement très faibles, de l'ordre de $10^{-3}$ lorsque plusieurs milliers de comptes ont le même centre d'intérêt.

Ainsi nos indicateurs permettent de mieux identifier et caractériser les groupes de comptes, et peuvent ensuite être combinés avec les mesures sur les individus introduites en chapitre 6 : au sein d'un groupe particulièrement pertinent, nous pouvons procéder à l'attribution de rôles et à l'identification d'individus-clés.

## 7.6 Synthèse

L'interaction entre comptes utilisateurs partageant les mêmes centres d'intérêt est propice au déclenchement de la propagation d'un message de manière plus large dans le réseau ; en conséquence nous voulons observer le réseau social à travers les groupes émergents qui le constituent.

La détection de communautés est réalisée à partir de différents modèles et graphes, via une variété d'algorithmes. Cependant, il manque à ces méthodes une phase de caractérisation, pour étudier la cohésion interne de chacun des groupes obtenus. Bien que des métriques topologiques existent dans l'état de l'art, elles ne prennent pas en compte le texte des messages échangés et ne permettent pas d'évaluer la cohésion thématique des communautés.

Dans ce chapitre, nous avons proposé un modèle en graphes pour représenter différents types de liens entre comptes utilisateurs (abonnements, interaction, partage), et proposons de marquer ces liens par leur thématique. Le cœur de notre contribution consiste en l'introduction de mesures de cohésion thématique, qui évaluent la relation des groupes envers le contenu des messages échangés. Combinées avec les mesures topologiques classiques, nos mesures permettent d'évaluer la force de cohésion interne d'un groupe, mais aussi d'estimer l'influence d'un groupe envers son environnement.

Cette contribution bénéficie des éléments extraits des textes émis et échangés, et place les comptes utilisateurs en relation les uns avec les autres, faisant apparaître des structures sociales mesurables. L'association de ces trois volets, analyse de texte, analyse des comptes, analyse des groupes, donne les clés de lecture sur trois niveaux.

Les travaux présentés dans ce chapitre ont fait l'objet de publications en conférences [Gadek et al., 2017b], dont une récompensée par l'obtention d'un *Best student paper award* [Gadek et al., 2017c], et d'un article de journal en cours de publication [Gadek et al., 2018].

# Troisième partie

# Évaluation du système

# CHAPITRE 8

# Le système SARTN

Les médias sociaux, où chacun a l'opportunité de devenir un émetteur médiatique, génèrent un important flux de messages qui peuvent être exploités à des fins publicitaires, de marketing, d'intelligence économique, ou encore en renseignement et sécurité. Cet espace informationnel soulève quelques défis : les messages échangés sont nombreux ; leurs auteurs sont mal identifiés ; des effets de groupe sont constatés.

Dans le but de répondre à notre problématique, nous avons construit et implémenté un système, nommé SARTN, pour *Système d'Analyse des Réseaux sociaux sur Trois Niveaux*, tirant profit des contributions théoriques introduites précédemment en analyse du sentiment et de l'opinion, en identification des acteurs-clés et caractérisation des comportements, et en analyse des graphes sociaux. Ces contributions permettent d'exploiter au mieux les messages, d'identifier les comptes d'utilisateurs, et d'explorer l'aspect social, via la détection et caractérisation des communautés.

Ces différents modules sont ici reliés en un système complet, dotant l'*analyste de réseaux sociaux* des outils nécessaires pour répondre à ses questions.

Une vue d'ensemble du système est proposée en section 8.1. La chaîne de traitement et ses trois niveaux d'analyse sont détaillés en section 8.2. La section 8.3 présente les technologies utilisées pour l'implémentation. Finalement, une synthèse conclut ce chapitre.

## 8.1 Vue d'ensemble des fonctionnalités

Le système proposé, SARTN, regroupe des fonctionnalités d'analyse portant sur trois niveaux : sur les textes, sur les comptes utilisateurs, et sur les groupes de comptes ; en figure 8.1, des couleurs différentes sont associées à ces niveaux d'analyse. Pour chacun de ces trois niveaux, les traitements réalisés sont de deux types : **en flux**, lorsqu'ils sont appliqués sur les messages reçus, un à un ; **en batch**, c'est-à-dire sur un jeu de données complet, lorsque c'est nécessaire.

La figure 8.1 montre ces fonctionnalités et informations calculées par le système, à partir d'un flux d'entrée : des messages, par exemple provenant de Twitter, qui sont constitués d'un *contenu* (par exemple, le texte), auxquels sont associées des informations à propos de l'*auteur*. Les traitements appliqués aux textes apparaissent en jaune sur la figure 8.1. Chaque message est traité indépendamment pour en déterminer le sentiment et le style d'écriture ; l'ensemble des messages est requis pour en extraire des thématiques, c'est-à-dire des *clusters* de documents. En parallèle, des informations d'interaction et de partage d'objets sociaux sont extraites du contenu des messages et apparaissent en bleu clair : ces informations sont nécessaires pour les comptes, et pour les groupes.

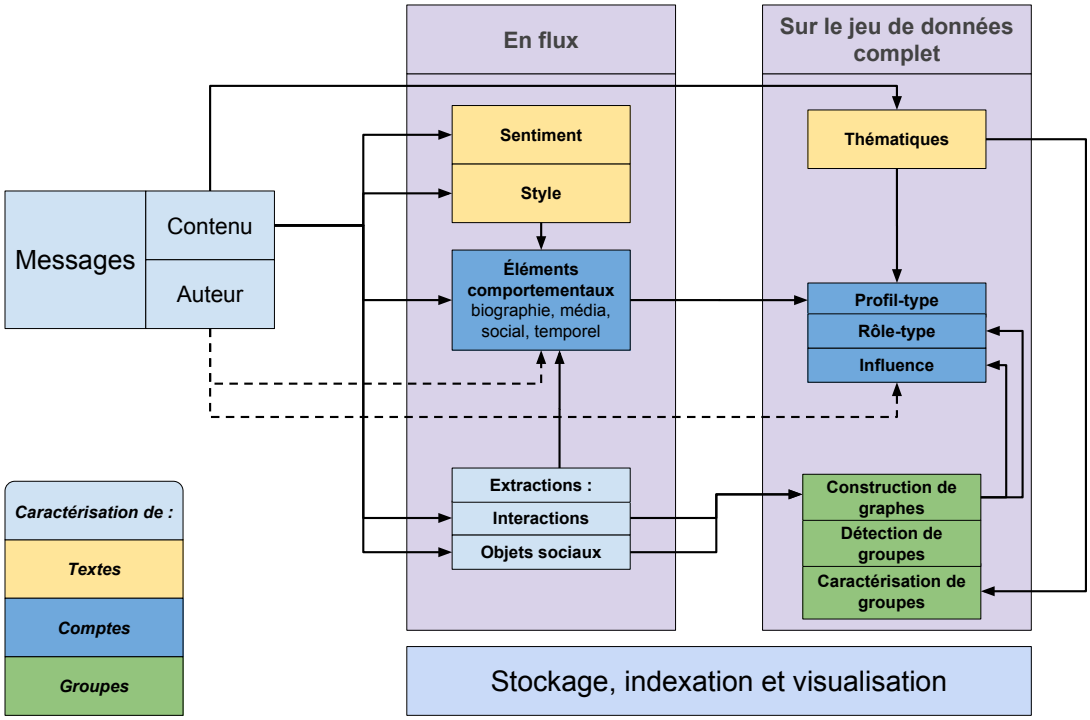



FIGURE 8.1 – Niveaux d'analyse et types de calcul

Les informations concernant l'auteur aident à caractériser le comportement ainsi que la popularité du compte (l'un des scores d'influence). Les flèches concernant ces informations figurent en pointillés sur le schéma, pour clarifier les croisements de flèches. La manière dont un compte utilise les interactions et objets sociaux vient rejoindre l'information de style d'écriture pour compléter les éléments comportementaux, puis les profils-types des comptes ; ces éléments apparaissent en bleu foncé.

L'interaction et le partage alimentent aussi l'analyse des relations et des groupes, en vert sur la figure 8.1, à commencer par la construction de graphes mettant les comptes en relation les uns avec les autres. Ces graphes sont nécessaires pour calculer le rôle-type des comptes ainsi que l'un des scores d'influence, mais surtout pour détecter des communautés d'utilisateurs. Enfin, les communautés sont évaluées d'une part sur leurs caractéristiques structurelles, topologiques (dans les graphes), mais aussi grâce à l'exploitation des thématiques issues des textes, en mesurant leur cohésion et pertinence sur un thème.

Finalement, tant les documents originaux que les éléments calculés sont stockés et indexés dans une base de données, ce qui facilite l'exploration rapide et la visualisation de tous les niveaux d'information par l'utilisateur de SARTN.

## 8.2 La chaîne de traitement, brique à brique

Dans cette section est décrite la chaîne de traitement, illustrée en figure 8.1. Dès réception des messages ou publications, le système extrait les interactions et objets sociaux, et procède à une analyse des textes, lors d'un traitement en flux. Le deuxième axe de traitement est focalisé sur l'étude des comptes : les informations sont agrégées sur une période de temps pour calculer les profils-types des utilisateurs, ainsi que leurs scores d'influence et leurs rôles-types. Le troisième axe de traitement construit des graphes contenant toutes les relations entre les comptes, précédemment extraites des messages collectés. Le système détecte alors les communautés à partir des graphes, et les évalue notamment au regard des informations de thématique.

Chacun de ces trois niveaux requiert tout d'abord des données, fournies par un collecteur spécifique au réseau social numérique visé.

### 8.2.1 Collecte des données

Le système SARTN est alimenté par un flux de données provenant d'une plate-forme sociale numérique ; il s'agit par exemple un flux de messages publiés sur Twitter. Il pourrait s'agir d'autres éléments, tels que des billets et des réponses provenant de Reddit. Dans tous les cas, les données reçues sont des objets (souvent représentés en *json*), qui se décomposent en deux parties :

— le **contenu** : souvent du texte, il peut aussi s'agir de liens, de photos, vidéos. Il s'accompagne systématiquement de **metadonnées** : la date et l'identifiant de l'auteur, parfois d'autres informations telles que le point d'émission.

— de l'information sur l'**auteur**. Au-delà d'un identifiant numérique (ID), d'autres informations sont souvent présentes et permettent de donner un relief au message : les textes de description biographique et de lieu de résidence (volontairement entrés et publiés par l'auteur), l'application utilisée pour publier, le nombre d'amis ou d'abonnés, etc.

Pour Twitter par exemple, l'obtention de ces deux ensembles d'informations est effectuée (ou complétée) via une API REST [1], permettant d'interroger spécifiquement le réseau social à propos d'un auteur, pour récupérer des informations non-jointes avec le message (selon l'implémentation de la plate-forme), ou pour actualiser des informations périmées.

Plus précisément, Twitter propose un accès gratuit limité à ses données ; via l'API *Stream*, il est possible de s'abonner à 5000 comptes (flux *user*), et d'en recevoir toutes les publications ainsi que tous les messages publics les mentionnant. Il est aussi possible de recevoir un échantillon de l'activité à propos de mots-clés (flux *keywords*, au maximum 500 mots), ou encore des messages géo-localisés (flux *geo*). Les flux *keywords* et *geo* ne garantissent absolument pas l'exhaustivité des messages.

D'autres plates-formes d'intérêt pour les utilisateurs de SARTN ne présentent parfois pas d'API de collecte. C'est le cas pour Galaxy2, petit réseau social dont une analyse est proposée en chapitre 9 : un contournement du problème consiste à collecter les pages *html* et à en extraire le contenu (faire du *scrapping*) pour reconstituer les publications, leur attribuer une date et un auteur.

### 8.2.2 Extraction d'informations : relations et contenus

Sur les messages reçus sont appliqués plusieurs processus d'extractions, qui enrichissent la donnée brute par des relations entre les comptes. Il s'agit de récupérer les interactions directes entre comptes, ainsi que les partages d'objets tels que les URLs ou les images.

#### 8.2.2.1 Données d'interaction

Les messages reçus, souvent émis en *broadcast* (volonté de publier, d'afficher, à tout destinataire potentiel), comportent toutefois des éléments d'interactions entre comptes d'utilisateurs. Une interaction est une action ponctuelle réalisée par un compte et impliquant un autre compte utilisateur, nous l'utilisons comme trace d'une proximité « sociale » entre les deux comptes. Il s'agit par exemple d'une citation, une réponse, un commentaire, un morceau de discussion entre deux comptes. Il peut aussi n'y avoir aucune trace d'interaction, ou seulement des traces implicites, qui sont alors invisibles pour le système. Le processus d'extraction des interactions consiste alors à parcourir le message à la recherche de mentions de comptes utilisateurs, mais aussi parfois d'autres motifs (RT pour retweet, « @ » précédant le pseudonyme sur Twitter, « /u/ » précédant le pseudonyme sur Reddit). Notons que certaines interactions, qui étaient faites « à la main » dans le passé, sont désormais intégrées dans les fonctionnalités des réseaux sociaux. Il s'agit notamment des réponses et des retweets, qui sont apparus comme une convention tacite entre utilisateurs, avant d'être transformés en une fonctionnalité de Twitter.

Lorsqu'elle correspond à une caractéristique propre du réseau social, l'interaction est plus aisée à récupérer : il suffit alors de recopier le champ correspondant dans le message reçu, en conservant

1. *Representational State Transfer*, un ensemble de contraintes définissant le format d'échange, basé sur HTTP.

le type de relation. En effet, une citation n'est pas une réponse, même si elles sont toutes deux des interactions.

#### 8.2.2.2 Données de partage d'objets

Les *objets sociaux* sont des contenus émis avec l'intention de les diffuser, de les faire circuler sur le réseau social, entre utilisateurs. Ils incluent notamment des URLs, hashtags, objets média tels que des photos, vidéos ou enregistrements audio, les mentions d'utilisateurs, fichiers partagés, etc. Un message peut n'en contenir aucun, ou plusieurs.

Comme ces objets sociaux sont émis et ré-émis par plusieurs comptes utilisateurs, ils permettent de mieux comprendre le partage et la diffusion de l'information sur le réseau. Afin de les tracer, un extracteur d'objets sociaux est nécessaire : il génère un inventaire des objets sociaux présents dans chaque message.

L'inventaire est stocké avec les informations suivantes :

— ID du message
— date de publication
— ID de l'auteur
— pour chaque type d'objets sociaux (URL, hashtag, média, etc) :
  — liste des URI des objets émis
  — stockage des objets média en local

Dans une extension possible, des traitements supplémentaires pourraient être appliqués sur les objets multimédia : par exemple, l'identification d'objets ou d'entités sur des photos donnerait lieu à l'ajout de labels. Ainsi, le contenu de l'objet média serait décrit et exploitable. Cette extension ne fait pas partie de nos travaux de thèse.

### 8.2.3 Analyse des données textuelles

Le système SARTN doit calculer l'information utile pour caractériser au mieux les discussions, les utilisateurs et les communautés. Chaque message est pré-traité : cette étape recouvre la détection de langue, la tokenisation (séparation de la phrase en mots), éventuellement, selon la langue, l'analyse grammaticale pour attribuer au mot un rôle (adjectif, nom, verbe, adverbe...). Les mots sans valeur sémantique (*stopwords*) peuvent être retirés. L'implémentation est spécifique au système, mais profite de la bibliothèque NLTK[2] *natural language toolkit*, en Python. Cette brique d'analyse des textes, illustrée en figure 8.2, se décompose en trois modules.

Un premier module correspond à l'implémentation du modèle théorique introduit en chapitre 6, dédié à la mesure du style d'écriture de l'auteur du message. Il est composé de plusieurs indicateurs : la longueur du message, la quantité de signes de ponctuation, de symboles, la quantité d'émojis (codées en UTF-8 ou mieux), d'émoticônes (par exemple, « :) ») et de hashtags. Ce module analyse chaque message dès qu'il est reçu, afin de l'indexer avec ces valeurs calculées.

Un second module détermine le sentiment du message dès sa réception ; il s'agit de l'analyseur *Vader* [Hutto and Gilbert, 2014], présenté en chapitre 2. À partir de ressources linguistiques, telle SentiWordNet [Baccianella et al., 2010] qui associe des termes et leur valeur en sentiment, ainsi que de quelques règles (par exemple pour gérer la négation), ce module détermine la polarité d'un message par un nombre réel, entre -1 et 1, qu'il est possible de remplacer par un label (positif, négatif ou neutre). Les textes neutres regroupent les textes objectifs, c'est-à-dire effectivement sans opinion, et les textes non-analysables. Notre choix a écarté le système de détection de posture avec contextualisation introduit en chapitre 5, car il nécessite trop de données étiquetées pour être générique. Nous y substituons le couple *(sentiment, thématique)* pour adresser le problème de l'identification

2. `http://www.nltk.org/`

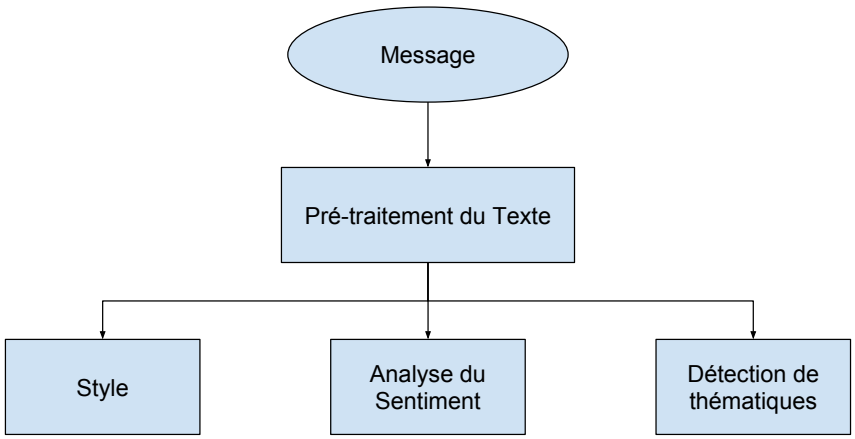



FIGURE 8.2 – Module de calcul du style

des opinions, malgré des limites inhérentes à l'approche : le sentiment n'est pas contextualisé, et la thématique ne remplace pas la subtilité de l'objet sur lequel porte l'opinion. Il s'agit d'un compromis : peu de données sont nécessaires, et les calculs sont rapides.

Enfin, le troisième module permet de détecter les thématiques dans un corpus, et d'étiqueter chaque texte avec sa thématique principale. Ce traitement nécessite donc d'être appliqué sur un gros volume de documents. Nous avons choisi une technique non supervisée, qui regroupe les textes en clusters de texte dont la projection en vecteurs *tf.idf* est similaire : Latent Dirichlet Allocation, abrégé LDA [Blei et al., 2003], réputée pertinente sur des textes courts tels que les tweets ou les billets de blogs. Cette approche non-supervisée offre ainsi une adaptabilité et une généricité qui répond aux besoins du système SARTN.

Ces trois modules couvrent un large spectre d'informations provenant du texte ; pour accroître cette couverture, plusieurs pistes sont envisagées. La première consiste à enrichir le style par la répartition des rôles grammaticaux des mots utilisés : quantité de pronoms, mode des verbes sont des indicateurs de la personnalité de l'auteur [Pennebaker et al., 2001]. Une seconde piste étend le module de sentiment, pour y ajouter la détection d'émotions. Enfin, bien que la détection non-supervisée de thématiques soit indépendante de la langue, ce n'est pas le cas du module d'analyse de sentiment. *Vader* est spécifique à l'anglais : un module équivalent est nécessaire pour prendre en compte une autre langue.

### 8.2.4 Analyse des comptes utilisateurs

L'analyse des comptes utilisateurs repose sur trois piliers, introduits en chapitre 6 : profil comportemental, indicateurs d'influence, et position sociale, ou rôle. La figure 8.3 illustre, à gauche, les données d'entrée, et produit, à droite, des informations de haut niveau exploitables par l'utilisateur de SARTN. La donnée exploitée provient des traitements appliqués sur les textes, des éléments comportementaux, et des données relationnelles qui, via des graphes dont nous ne détaillons la construction que plus tard, permettent de calculer des rôles et un indicateur d'influence.

Premier module de l'étude des comptes, les **profils** des comptes utilisateurs sont représentés par des données réparties en cinq aspects, afin de caractériser les différentes facettes du comportement. Dans un second temps, les profils de toute la population du jeu de données sont exploités pour en dégager des profils-types, facilitant leur lecture par l'utilisateur du système. Les cinq aspects retenus sont les suivants :

— Biographie : l'identité de l'utilisateur. Nom, pseudonyme, ID, champs textuels libres, date de

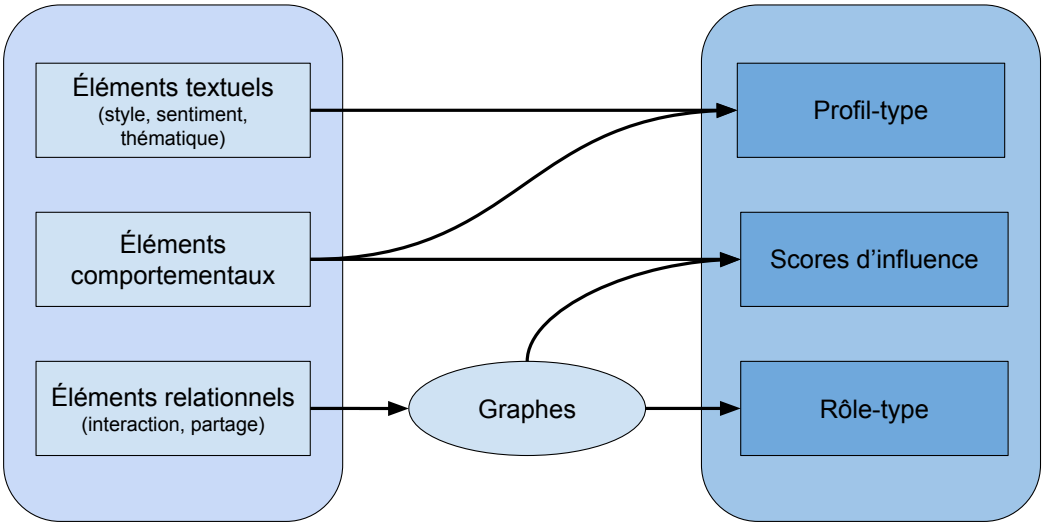



FIGURE 8.3 – Vue générale de l'analyse des comptes utilisateurs

création, fuseau horaire, langue de l'interface, application d'émission des messages ;

— Style : longueur du message, quantité de signes de ponctuation, de symboles, quantité d'émojis, d'émoticônes, de hashtags. Thèmes et sentiments des messages ;

— Média : nombre total et moyen quotidien d'objets média émis ;

— Interaction : pour chaque type d'interaction, nombres total et moyen d'interactions émises. Nombre d'utilisateurs différents mentionnés ;

— Temporel : histogramme de l'activité du compte, temps moyen entre deux messages.

Les profils sont stockés séparément de la base principale de stockage des documents reçus. En effet, il s'agit ici de données numériques : il est utile de les récupérer rapidement, soit par individu, soit par date, pour les fournir à des outils de calcul ou de visualisation statistique. À partir de ces profils, nous calculons des profil-types, selon la méthode décrite en chapitre 6.

En parallèle, le graphe des interactions, $G_I$, ainsi que les statistiques d'engagement (nombre d'abonnés, nombre de rediffusions ou de commentaires suscités) des messages émis alimentent les **scores d'influence** : les scores de référence et d'expertise sont déterminés en *batch*, nécessitant l'exploitation du jeu de données complet ; le score de popularité est calculé message après message, et son évolution est parfois critique (il y a parfois achat d'abonnés, résultant en une forte augmentation instantanée de la popularité).

Enfin, le calcul du **rôle**, c'est-à-dire du type de position occupée par les utilisateurs dans un graphe, nécessite tout d'abord la construction du graphe d'interaction (explicité dans la section suivante) ; ce calcul est donc réalisé en *batch* sur le jeu de données complet. La méthode retenue est RolX [Henderson et al., 2012], dont une implémentation compatible avec *NetworkX* est disponible en source ouverte [3] : il s'agit de la méthode la plus complète disponible, qui accomplit à la fois la tâche de calculer les caractéristiques topologiques, ainsi que les tâches de clustering et de « *sense making* », un résumé des degrés et centralités typiques, par rôle.

Ces trois types d'indicateurs couvrent plusieurs dimensions (comportement, engagement et popularité, position dans le graphe) pour caractériser les comptes utilisateurs. Cependant, le comportement social de ces comptes fait émerger des groupes d'individus, que l'analyste souhaite détecter et caractériser.

3. `https://github.com/Lab41/Circulo/blob/master/circulo/algorithms/rolx.py`

### 8.2.5 Construction de graphes et obtention de groupes d'influence

Une fois collectée et extraite, les informations relationnelles (abonnements, interactions, partage d'objets sociaux) sont traitées de manière à obtenir des *groupes d'influences*, groupes d'utilisateurs fortement liés tant par l'interaction que par les thématiques abordées. La figure 8.4 représente la partie de la chaîne de traitement réalisant la construction de différents graphes reliant les utilisateurs entre eux, et l'application de traitements (filtrage et partitionnement) débouchant sur la détection et la caractérisation de communautés d'utilisateurs.

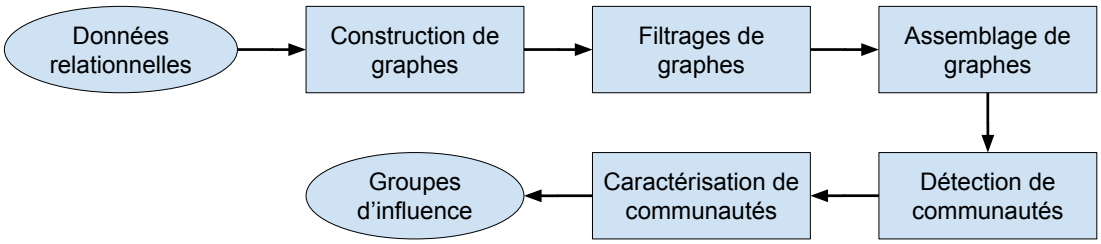



FIGURE 8.4 – Analyse de graphes et obtention des groupes d'influence

Les **données d'entrée** sont constituées par les données d'interaction directe entre utilisateurs (abonnement, citation, mention, publication sur une page), ainsi que par les données de partage d'objets sociaux (publication simultanée d'images, de photos, de hashtags). Lorsque la donnée relationnelle est portée par un message textuel, une thématique l'accompagne et sert à la construction d'un graphe d'interaction thématique.

La **construction** produit plusieurs graphes, introduits en chapitre 7 : d'amitié $G_F$, d'interaction $G_I$, d'interaction thématique $G_I^\theta$, et de partage, $G_\Omega$. Les graphes manipulés par le système sont dirigés et leurs arcs sont valués. Par exemple, $G_I$ est un graphe dirigé : pour chaque interaction détectée dans la période T d'intérêt, un arc dirigé valué est ajouté, entre un premier utilisateur, auteur de l'interaction, et sa destination. Un message portant $n$ interactions donne lieu à la création (ou l'ajout d'un poids de valeur 1) de $n$ arcs, de l'auteur vers chacun des destinataires.

Un mécanisme de **filtrage** est disponible pour retirer des arcs ou des nœuds. Les arcs peuvent être retirés selon le poids qui y est associé : des arcs trop faibles, par exemple de valeur 1, ne représentent probablement pas de grand lien entre les deux comptes. Ils peuvent aussi être retirés selon la réciprocité du lien, en ne considérant les arcs $e_{i,j}$ qu'à la condition que $e_{j,i}$ soit non nul ; ou bien encore en considérant les dates d'interaction, via une fonction attribuant le poids de l'arc en fonction des dates de réalisation de l'interaction. Par défaut, le graphe est conservé tel quel ; l'utilisateur de notre système dispose des outils selon son besoin.

Ces traitements sont spécifiques aux types de graphes. Ainsi, les abonnements sont très majoritairement uniques et ponctuels : aucun seuillage n'est requis. Sur le graphe d'interaction, il s'agit d'un compromis : si le graphe est trop gros, l'importance d'actions ponctuelles et uniques (par exemple, l'unique *retweet* de la part d'un compte) est affaiblie, et l'utilisateur de SARTN préférera éliminer les nombreux nœuds qui présentent un unique arc, allégeant l'empreinte mémoire et les temps de calcul.

La phase d'**assemblage** suit le même principe : l'union, l'intersection ou encore l'analyse séparée sont disponibles. Notre recommandation *a priori* consiste à exploiter l'intégralité du graphe des interactions, qui constitue une première approche pertinente.

Sur le ou les graphe(s) ainsi obtenu(s), nous appliquons un algorithme de **détection de communautés**. Informatiquement, une communauté est un ensemble d'individus, c'est-à-dire une liste de

nœuds. Notre choix s'est arrêté sur une méthode de partitionnement, l'algorithme *Louvain* [Blondel et al., 2008], qui n'autorise pas l'appartenance d'un nœud à plusieurs groupes. D'autres algorithmes similaires, donnant des résultats souvent proches mais différents, sont *InfoMap*, *Walktrap* et *Fast-Greedy*; une comparaison des ces méthodes est proposée en chapitre 4.

Au-delà des partitions, une autre piste semble pertinente : les couvertures, c'est-à-dire la détection de communautés autorisant l'appartenance multiple d'un nœud à plusieurs groupes. Cependant, ces méthodes sont plus difficiles à valider, et certaines d'entre elles affichent des complexités exponentielles. Nous en fournissons une revue en chapitre 4.

L'implémentation de Louvain retenue est disponible en tant que bibliothèque Python : il s'agit du module *python-louvain* [4].

À la phase de détection succède une phase de **caractérisation des communautés**. Comme nous l'avons présenté dans le chapitre 7, le système SARTN analyse les communautés en générant des indicateurs de qualité, répartis en deux catégories : indicateurs topologiques, basés sur le graphe ; et indicateurs thématiques, nécessitant l'analyse des documents originaux.

TABLEAU 8.1 – Indicateurs de qualité des communautés

| Type | Topologiques | Thématiques |
|---|---|---|
| Mesures | $d_{int}$, TPR, c, taille | $\xi_u$, $\xi_t$, $\rho_u$, $\rho_t$ $\xi_{sim}$, $\theta f.igf$ |

La densité interne ($d_{int}$) évalue la quantité de liens au sein d'un groupe ; la proportion de triangles TPR mesure la quantité de membres isolés (ou mal reliés au groupe). La conductance *c* évalue les liens entre la communauté et son environnement. Enfin, la taille de la communauté correspond au nombre de ses membres.

L'expertise ($\xi_u$, $\xi_t$, $\xi_{sim}$) mesure la proportion d'utilisateurs ou de messages de même thématique au sein d'un groupe. La représentativité ($\rho_u$, $\rho_t$) évalue l'importance d'un groupe sur une thématique, et $\theta f.igf$ sa pertinence.

Ces indicateurs signalent rapidement les groupes présentant une forte cohésion structurelle, une forte croissance en taille, et/ou une excellente cohésion sémantique. Ces groupes sont supposément des *groupes d'influence*, leurs bons scores étant les symptômes de l'impact de leurs messages.

L'utilisateur de SARTN peut également restreindre dans un premier temps le champ de ses recherches à un thème (par exemple, issu de l'algorithme LDA) ou bien à un ensemble de mots-clés, sur lequel il souhaite voir comment s'organise le paysage social, et les influences de chaque groupe.

Enfin, ces groupes d'influence prennent aussi leur sens en visualisant simplement les thèmes, textes et objets sociaux qui sont le plus échangés en leur sein. Le stockage de ces objets sociaux et les liens conservés envers leurs auteurs rendent cette visualisation rapide.

## 8.3 Détails d'implémentation technique

Nous présentons en figure 8.5 l'ensemble des composants technologiques dont dépend le système SARTN. Ces composants sont répartis sur trois niveaux : d'infrastructure (gestion du matériel) et stockage (index pour stoker les documents), de traitement (outils de codage et de calcul), et de visualisation. Chacun de ces composants est présenté ci-après.

**Hadoop** [5] est un *framework* dédié à l'exploitation d'un ensemble de machines. Il sert de support au calcul distribué et à la scalabilité, notamment par son système de fichiers HDFS qui fait abstraction de la machine sur laquelle un fichier est stocké (les fichiers sont distribués et répliqués de manière transparente pour le développeur). Il inclut une implémentation de *MapReduce*, une méthode de parallélisation de tâches en deux étapes de calculs unitaires.

4. https://python-louvain.readthedocs.io/
5. https://hadoop.apache.org/

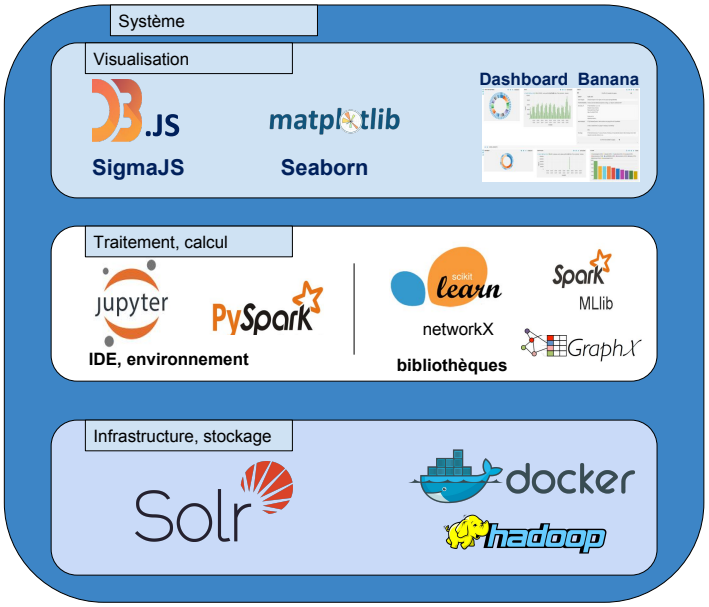



FIGURE 8.5 – Système global et technologies utilisées

**Docker** [6] est un outil de virtualisation : chaque module applicatif (par exemple, la base de données) est isolé dans un *container* (conteneur). Tous les conteneurs partageant le même système d'exploitation, cela diminue leur taille, et facilite l'intégration de briques supplémentaires ainsi que la montée en charge dynamique.

Le moteur d'indexation **SolR** [7] est capable d'ingérer des millions de documents, et de répondre rapidement aux requêtes textuelles. Bien que disposant d'un mode « sans schéma » qui lui permet d'ingérer des documents *Json* sans a priori sur leur structure, la rapidité (d'indexation et de requête) est nettement améliorée lorsqu'un schéma est spécifié.

Les *notebooks* proposent une approche interactive pour coder, où une session est maintenue active, permettant d'exécuter quelques lignes de code à la fois. **Jupyter** [8] permet ainsi de dérouler tout le traitement, et d'explorer les résultats, en générant des graphiques visibles dans le « cahier ». Cet environnement de développement est propice au prototypage : l'analyse des résultats intermédiaires permet de prendre des décisions lors du traitement, et d'appliquer au plus tôt des correctifs nécessaires.

Sur une infrastructure basée sur *Hadoop*, il est courant d'utiliser **Spark**, qui mutualise les ressources de calcul (RAM et CPU) et rend aisée l'utilisation de *MapReduce*. Les langages de programmation disponibles incluent Java, Scala ou Python, ce dernier via l'interface *pyspark*.

Les principales bibliothèques utilisées sont **Scikit-learn** [9], dédiée à l'apprentissage automatique, **NetworkX** [10], qui fournit de nombreux outils de manipulation et de calcul de graphes, ainsi que leurs *alter ego* dans l'univers Spark : SparkML et GraphX, moins complètes mais plus rapides sur de grands jeux de données.

Enfin, la visualisation repose sur trois pans : le premier consiste en un tableau de bord, appelé **Banana** [11], qui affiche les contenus de l'index de documents SolR sur une vue interactive grâce à

6. https://www.docker.com/
7. https://lucene.apache.org/solr/
8. https://jupyter.org/
9. http://scikit-learn.org/
10. http://networkx.github.io/
11. https://github.com/LucidWorks/banana

des diagrammes, courbes temporelles et affichage de documents. Le second pan est fourni par les bibliothèques classiques de Python, **MatplotLib** [12] et **Seaborn** [13], qui génèrent des courbes et des nuages de points. Enfin, les bibliothèques javascript **D3js** [14] et **sigmajs** [15] fournissent des outils légers de visualisation de données pour *D3*, et de manipulation interactive de graphes pour *sigma*.

## 8.4 Synthèse

La quantité et diversité des messages échangés sur un réseau social numérique pose un vrai défi, scientifique et technique. Les traitements sur les données que nous proposons dans les chapitres 5, 6 et 7 répondent scientifiquement à ce problème.

Dans ce chapitre, nous avons présenté SARTN, un système qui tire profit des technologies de distribution des calculs pour plus facilement traiter des volumes de données importants ; cette implémentation donne la capacité de réaliser une analyse des textes, des comptes et des groupes issus des graphes sociaux, en recourant à des traitements en flux et en batch, quand c'est nécessaire. Dans le cadre d'un projet de recherche interne à Airbus, le système SARTN a été déployé sur une infrastructure « cloud » privée, au *DataLab*, hébergée sur le site de Toulouse. Ce système a donné lieu au dépôt d'un brevet, auprès de l'Office Européen des Brevets (OEB), en octobre 2017.

Dans le chapitre suivant, nous nous penchons sur deux jeux de données issus de deux réseaux sociaux, Twitter et Galaxy2, afin d'évaluer la pertinence du système proposé. Ces deux études de cas exploitent les contributions de cette thèse pour identifier les comptes et groupes importants, dans un premier temps à propos de politique américaine, et dans un second temps sur le *Darknet*.

12. `https://matplotlib.org/`
13. `https://seaborn.pydata.org/`
14. `https://d3js.org/`
15. `http://sigmajs.org/`

# CHAPITRE 9

# Étude de cas sur deux réseaux sociaux numériques

Le système SARTN précédemment décrit est exploité sur deux exemples d'application, alimentés par des données réelles. La première étude est basée sur un corpus de plusieurs millions de tweets, concernant la campagne électorale américaine de 2016. La seconde étude est fondée sur l'intégralité des messages émis sur Galaxy2 ; de taille modeste, ce corpus nous permet de décliner notre système sur un autre média social, pour vérifier sa généricité et son adaptabilité.

Afin d'évaluer la pertinence de nos contributions, nous introduisons des critères fonctionnels concernant les trois niveaux d'analyse des réseaux sociaux : sur les textes, les comptes utilisateurs, et les communautés. Durant les deux études, nous nous y référerons.

Les critères d'évaluation du système sont introduits en section 9.1. L'étude du corpus de tweets est présentée en section 9.2 ; puis nous nous penchons sur Galaxy2 en section 9.3. Enfin, la dernière section récapitule les critères atteints et les limites du système proposé.

## 9.1 Critères d'évaluation

Afin d'estimer la qualité de notre réponse à la problématique-titre, *détection d'opinions, d'acteurs-clés et de communautés thématiques*, nous introduisons des critères d'évaluation du système, pour accompagner les deux études de cas.

**Critère 1.** *Le système évalue les opinions présentées sur le réseau social.*

Le critère 1 couvre l'analyse des messages : l'utilisateur a accès aux contenus originaux, mais doit aussi disposer d'éléments pour quantifier les opinions, et trier les messages par thématiques et sentiments.

**Critère 2.** *Le système identifie des acteurs-clés du réseau social et permet de comparer les influences de différents comptes.*

Face à la quantité des comptes produisant des contenus sur les médias sociaux, le système doit identifier des comptes importants dans le réseau, facilitant son exploration. Des indicateurs doivent permettre d'estimer et de comparer l'importance de comptes, comme décrit dans le critère 2.

**Critère 3.** *Le profil et le rôle fournissent des informations complémentaires pour caractériser les comptes.*

Profil et rôle sont des notions introduites dans le chapitre 6, caractérisant respectivement le type de comportement et la position sociale d'un compte utilisateur ; le critère 3 s'assure que ces notions sont différentes et complémentaires.

**Critère 4.** *Le système exploite différents types de relations liant les comptes.*

Les réseaux sociaux numériques fournissent aux individus différentes manières de tisser des liens entre eux : abonnements, discussions, partages, et parfois des fonctionnalités spécifiques aux plate-formes. Le critère 4 vérifie que ces relations sont exploitables, qu'il est possible de parcourir les comptes de proche en proche et de comparer les usages faits de ces fonctionnalités.

**Critère 5.** *Le système détecte les structures sociales émergentes et les caractérise.*

À partir des relations sociales entre comptes, le critère 5 évalue la capacité du système à détecter des structures sociales ou communautés : des groupes de comptes qui échangent beaucoup entre eux, et peu avec le reste du réseau. Des outils et mesures doivent permettre à l'utilisateur du système d'explorer efficacement ces groupes détectés.

**Critère 6.** *Les mesures thématiques sont plus informatives que les mesures topologiques introduites précédemment pour caractériser les communautés détectées.*

Parmi les mesures de qualité des communautés détectées précédemment citées, le système implémente des mesures thématiques, issues de notre contribution théorique en chapitre 7. Le critère 6 estime leur pertinence.

Nous faisons référence à ces critères dans les deux études de cas réalisées grâce à notre système, la première sur des tweets, et la seconde sur Galaxy2. Il nous paraît nécessaire de prouver que SARTN n'est pas dépendant des fonctionnalités originales proposées par une plate-forme (par exemple, les *retweets*), mais parvient tout de même à les exploiter.

## 9.2 Application à Twitter

La première de nos études de cas porte sur un corpus de tweets, émis à l'occasion de la campagne électorale américaine de 2016, qui s'est conclue par l'élection de Donald Trump à la présidence. Dans cette section, nous décrivons le corpus, puis nous exposons les analyses réalisées par notre système, abordant les messages, les comptes utilisateurs et les groupes de comptes.

### 9.2.1 Description du corpus *KevRandTweets*

Nous avons collecté le corpus *KevRandTweets*, composé de 9 671 711 tweets rédigés en anglais, via l'API de *Stream* proposée par Twitter. Cette API donne l'accès aux messages, dès leur publication, selon trois méthodes de requête : par auteur, par mot-clé ou par aire géographique (auquel cas, seuls des messages géolocalisés seront reçus). L'accès est restreint en quantité : un maximum de 5 000 comptes et 500 mots-clés (pour ces derniers, l'exhaustivité du flux n'est pas garantie) limite la collecte gratuite. Afin d'obtenir des relations entre les comptes, il fut choisi de réaliser une collecte par les 5 000 comptes, en sélectionnant ces comptes de façon à observer des contacts entre eux : ils sont tous « amis » d'un même compte.

Ce jeu de données présente quelques biais : il contient des comptes mentionnés, mais inactifs ; pour de nombreux comptes (qui ne font pas partie de la liste des 5 000 comptes-graines), nous ne disposons pas de l'intégralité des actions accomplies. La sélection des sources de la collecte, bien qu'argumentée, reste arbitraire. De plus, le corpus est bien évidemment limité aux tweets publics, excluant de fait les échanges par d'autres interfaces (notamment les blogs et les autres plate-formes) ou privés (par message direct par exemple). Cependant, l'API à notre disposition, c'est-à-dire la source de données elle-même, ne permet pas d'obtenir un flux complet ou d'isoler parfaitement un sous-graphe connexe du réseau social : notre système d'analyse de réseaux sociaux se satisfera de cette situation.

### 9.2.2 Analyse du texte : répartitions entre thématiques et sentiments

L'association d'un sentiment à chaque document est une première étape ; pour en tirer profit à grande échelle, le *Sentiment Net* donne une unique valeur pour un ensemble de documents. En posant *#Positifs* (resp. *#Negatifs*) la quantité de messages de sentiment positif (resp. négatif) dans une sélection donnée, il s'agit tout simplement du ratio suivant : $Net = \frac{\#Positifs - \#Negatifs}{1+\#Negatifs+\#Positifs}$. Cela permet de rapidement mettre en relief la fréquence de mots-clés ou de hashtags, comme nous le montrons en figure 9.1. Le corpus est composé de messages pro-Trump, ce qui favorise les termes #Donald, #Trump et #MAGA (l'acronyme du slogan, *Make America Great Again*), et place négativement l'ancien président Obama, ainsi que les *FakeNews*. Plus précisément, #MAGA est mentionné dans plus de 160 000 tweets dans le corpus, mais ces occurrences sont relativement équilibrées entre les sentiments positifs et négatifs, résultant en une moyenne autour de 0.1. Une visualisation du sentiment par auteur fait écho à cette vue par contenu, et est présentée lors de l'étude de Galaxy2, en figure 9.26.

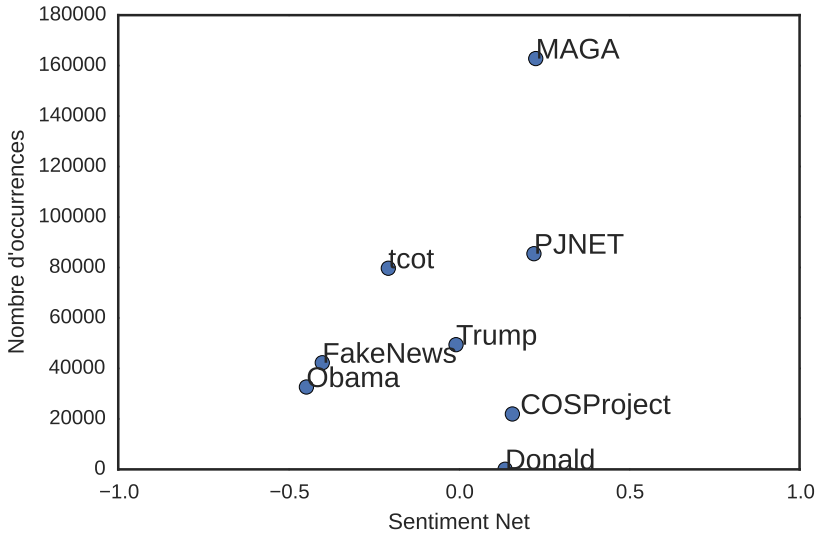



FIGURE 9.1 – Sentiment Net pour les hashtags les plus fréquents dans *KevRandTweets*

Une seconde étape d'analyse du texte repose sur la détection automatique de thématiques, permettant de rapprocher des documents similaires. L'approche retenue par SARTN, décrite dans le chapitre précédent, consiste en du *clustering* de textes via la méthode *Latent Dirichlet Allocation* [Blei et al., 2003]. Cette approche requiert un nombre de thématiques en paramètre *a priori*, que nous fixons à 40 : il s'agit d'un compromis entre la complexité du calcul, la focalisation thématique naturelle du corpus (centré sur la politique post-campagne électorale), la taille des documents et leur nombre. Des expériences avec un nombre de thèmes plus grand n'ont pas été probantes. Un outil de visualisation des thématiques accompagne cette annotation ; pour éviter les répétitions, il est décrit lors de la seconde étude de cas, en figure 9.24.

L'analyse de ce jeu de données nécessite que le système fournisse une visualisation. Pour cela, les tableaux de bord (ou *dashboards*) tels que Banana [1] mettent à disposition des outils, parfois biaisés, mais difficilement contournables. Des diagrammes répartissent les documents par auteurs, par voie d'émission, ou encore par thématiques, chacune découpée en trois sentiments (positif, neutre, négatif) ; une vue temporelle positionne les documents dans le temps. Les documents sont accessibles directement dans un tableau. L'ensemble de ces visuels interactifs reflètent une même sélection de données ; il est ainsi très simple de créer un filtre par hashtag, puis par période de temps, diminuant

1. Documentation accessible sur : `https://doc.lucidworks.com/lucidworks-hdpsearch/2.5/Guide-Banana.html`

de cette façon la quantité de documents concernés.

La figure 9.2 donne un aperçu de notre configuration du tableau de bord. Le premier diagramme montre les thématiques et sentiment : en un clic, l'utilisateur restreint la vue à une thématique donnée, ce qui actualise tout le tableau de bord. La répartition temporelle (second diagramme) est notamment mise à jour. L'utilisateur choisit alors une période de temps plus précise : par exemple, un pic d'activité particulièrement inhabituelle. Il voit sur la droite un tableau simplifié des données, contenant uniquement les pseudonymes et les textes des messages (les IDs sont cachés mais accessibles en deux clics). D'autres diagrammes répondent à des questions plus spécifiques : à mi-hauteur, sur la gauche, apparaît uniquement la répartition en sentiment (s'affranchissant des thématiques) ; au milieu, la quantité de messages publiés par les comptes concernés depuis leur création ; à droite, la quantité de messages répartis selon leurs auteurs, parmi la sélection courante. La dernière ligne contient d'autres diagrammes, qui listent les objets sociaux partagés : images, URLs, hashtags. Hors champ, d'autres modules proposent de faire des requêtes par mot-clé, par auteur, ou selon n'importe quelle caractéristique indexée ; des histogrammes montrent notamment des nuages de mots-clés, les applications utilisées (iPhone, Android, Web...), et les utilisateurs les plus mentionnés.

La visualisation par tableau de bord, en combinant des éléments reconnus dans les tweets (texte, hashtags, dates d'émission) et des éléments calculés (le sentiment et la thématique), donne toutes les clés pour que l'utilisateur du système prenne connaissance de la teneur des débats en cours sur le réseau social. Cette vue « sentimentale » répond au besoin énoncé en critère 1, mais il ne s'agit que d'un compromis : l'opinion, qui nécessite une identification fine du sujet du texte, est beaucoup plus précise et permet une analyse plus poussée ; cependant il est encore difficile d'identifier précisément l'objet de l'opinion et surtout, de combiner les opinions de textes mentionnant plusieurs aspects d'une même entité. L'approche par sentiment résulte globalement en de meilleurs taux d'erreurs, mesurés par le $F_1 - score$ ; nous avons décrit ce phénomène en chapitre 2.

La contribution introduite en chapitre 5, concernant la désambiguïsation par la contextonymie pour la détection de posture, n'est pas implémentée par le système SARTN. La posture est certes pertinente pour qualifier les débats sur un média social, mais le vocabulaire spécifique rend nécessaire un corpus d'apprentissage dédié à l'étude de cas, aussi le système exploite-t-il deux méthodes plus génériques : le sentiment et la thématique.

### 9.2.3 Représentation du réseau social par des graphes

Le système SARTN exploite des graphes pour les deux niveaux d'analyse suivants, concernant les comptes utilisateurs, puis les communautés thématiques. Cette section introduit les différents graphes exploités dans l'étude du corpus *KevRandTweets*.

Nous avons défini quatre graphes en chapitre 7, qui correspondent à des applications et exploitations différentes. Dans le cas de l'étude de *KevRandTweets*, les relations d'abonnements ne sont pas disponibles, ce qui empêche la construction du graphe « **des amitiés** », noté $G_F$.

Le graphe des **interactions** $G_I$ représente l'activité réalisée sur le réseau social. Chaque retweet, mention et réponse contribue d'un poids unitaire à l'arc dirigé de l'émetteur de l'action, vers le compte cible (qu'il soit retweeté, mentionné ou répondu).

Il est possible de décliner le graphe précédent en **interactions par thématique** : les graphes $G_I^\theta$ s'obtiennent en ne considérant que les messages de même étiquette $\theta$, résultant en quarante graphes (le nombre de thématiques retenu précédemment, lors de la phase d'exploitation des contenus textuels). À titre d'exemple, nous proposons de détecter des communautés sur chacun de ces graphes, puis de créer le graphe thématique $G_\Theta$ : lorsque, pour une thématique donnée, deux comptes sont présents dans une même communauté, alors l'arc entre eux est incrémenté de 1. Une étape de filtrage est nécessaire pour retirer les arcs de poids trop faible, avant de détecter les communautés sur $G_\Theta$.

Enfin, il est souvent intéressant de relier les utilisateurs selon les **objets sociaux** (hashtags, URLs, images) qu'ils partagent : deux comptes mentionnant le même site Web présentent une certaine proxi-


FIGURE 9.2 – Extrait du tableau de bord

mité. En reliant le compte émetteur $u \in U$ à l'objet social $\omega \in \Omega$ émis, le graphe bipartite des partages $G_\Omega = (U, \Omega, E)$ est construit, et alimente également la méthode de détection des communautés.

Afin de pouvoir comparer les trois partitions obtenues (respectivement notées $\Gamma_I$, $\Gamma_\Theta$, $\Gamma_\Omega$), nous calculons les modularités sur le graphe $G_I$. En effet, il n'est pas pertinent de comparer des modularités calculées sur des graphes différents ; de plus, les actions visibles telles que les retweets et les mentions nous semblent un support plus intuitif pour évaluer des communautés.

TABLEAU 9.1 – Comparaison de trois graphes

| Graphe | Modularité Q | Taille médiane | Nombre de groupes d'au moins 4 membres |
|---|---|---|---|
| $G_I$ | 0.446 | 14 | 82 |
| $G_\Theta$ | 0.097 | 6 | 77 |
| $G_\Omega$ | 0.406 | 379 | 87 |

La table 9.1 regroupe les valeurs de modularité, la taille médiane (en nombre de comptes par groupe) des communautés détectées, et le nombre de communautés suffisamment grandes (à partir de quatre membres). Au graphe $G_\Theta$ correspond une très faible modularité : les groupes détectés ne sont pas pertinents au regard des interactions globales. Pour le graphe des partages d'objets, la taille médiane des groupes est assez élevée : de nombreux hashtags et URLs sont très populaires et créent de trop nombreux liens, résultant en un découpage en peu de communautés. Le graphe des interactions n'est pas exempt de critiques : la modularité de sa partition reste faible, et la taille médiane cache des disparités, certains groupes contenant plusieurs milliers de nœuds. Nous choisissons alors de conserver $G_I$ et de réaliser la suite de l'analyse sur le support des interactions directes entre comptes.

Sur un jeu de données modeste tel que *KevRandTweets*, $G_I$ constitue déjà un défi avec 734 888 nœuds et 5 220 821 d'arcs dirigés et valués. Afin d'aborder ce graphe, nous nous intéressons aux communautés ou groupes de comptes, qui sont détectées en utilisant l'algorithme Louvain. Il faut ensuite les décrire et les caractériser. Dans un premier temps, cette description recourt aux informations de profil-type calculées dans la section précédente ; dans un second temps, les groupes sont caractérisés en exploitant les thématiques détectées dans chacun des tweets du corpus.

### 9.2.4 Comptes influents et types de comportement

Après l'aperçu fourni par l'étude des textes, nous nous intéressons désormais à l'activité des comptes utilisateurs. Dans un premier temps, nous sommes à la recherche des acteurs-clés du jeu de données, comparant les statistiques brutes (nombre d'abonnés, de tweets) et les scores d'influence introduits en chapitre 6. Dans un second temps, nous caractériserons les comportements des utilisateurs du réseau social numérique analysé.

#### 9.2.4.1 Détection des acteurs-clés du réseau

L'identification des acteurs-clés pose la question suivante : qui sont les comptes les plus importants dans la vue du réseau social à notre disposition ? La réponse est multiple, car l'importance des comptes recouvre plusieurs dimensions.

Une première approche consiste à classer les comptes en fonction du nombre de tweets émis, présents dans le corpus ; dans le cas de *KevRandTweets*, il s'agit de comptes tels que *marshawright* (plus gros émetteur du corpus) ou *usfreedomarmy*, dont les 7 566 tweets sont équitablement répartis quotidiennement, pour une moyenne de 10 messages par heure. Cet indicateur n'est donc pas très pertinent, puisqu'il met en avant des comportements de spam. Le constat fait en chapitre 6, écartant l'activité des indicateurs de l'influence, est ici validé. Certes, il faut un minimum d'activité pour être influent ; cependant les comptes les plus actifs sont fréquemment du spam, c'est-à-dire du bruit.

Une seconde approche consiste à mesurer les comptes par le nombre d'abonnés (*followers*). La table 9.2 liste les 10 comptes les plus populaires, ayant émis au moins un tweet recueilli dans le corpus. Le nombre d'abonnés est donné, ainsi que sa conversion en score de **popularité** (logarithme en base *MAX_FOLLOWERS*, qui répartit les scores entre 0 et 1), que nous avons introduite en chapitre 6. Arrivent en tête les comptes de Google et de Victoria's Secret, dont les services après-vente répondent systématiquement aux remarques et demandes des clients les mentionnant ; un comportement similaire explique la présence de Windows et d'Ubisoft. Le score de popularité permet d'apprécier plus facilement la taille des audiences ; le paramètre retenu attribue un score de 1 aux comptes ayant cent millions d'abonnés. Ces valeurs ont été calculées lors de l'émission des tweets, dans la période analysée.

TABLEAU 9.2 – Top10 des comptes selon la popularité

| Nom | Abonnés | Popularité |
|---|---:|---:|
| Google | 16 513 689 | 0.902 |
| VictoriasSecret | 10 351 953 | 0.877 |
| johnlegend | 8 444 059 | 0.866 |
| guardian | 6 109 452 | 0.848 |
| Windows | 5 967 413 | 0.847 |
| WalkingDead_AMC | 5 288 987 | 0.84 |
| Ubisoft | 5 086 466 | 0.838 |
| Oceaanfietser | 5 042 692 | 0.838 |
| dumbassgenius | 4 319 974 | 0.829 |
| wikileaks | 4 145 384 | 0.827 |

Cette popularité indique qu'un compte est connu, mais ne présage pas de sa capacité à susciter l'engagement. De plus, la popularité est fixe, quelle que soit la couverture du jeu de données : elle ne dénote pas l'importance sociale du compte au sein du sous-ensemble du réseau social numérique contenu dans le jeu de données analysé. Pour pallier ce problème, l'indicateur d'**influence**, que nous avons introduit en chapitre 6, classe les nœuds (les comptes utilisateurs) d'un graphe construit sur les interactions (c'est-à-dire les mentions, les retweets et les réponses sur Twitter). La table 9.3 présente les dix comptes les plus influents selon ce score, dans l'ordre, et donne à titre comparatif le nombre d'abonnés, ainsi que les scores de **popularité** et d'**expertise**.

Le score d'**expertise** est calculé selon la quantité de retweets concernant les messages dont un utilisateur est l'auteur, et prend en compte la thématique des tweets émis. L'échelle retenue est logarithmique : il faut toujours plus de retweets pour progresser le long de ce score ; pour obtenir un score de 1, un total de 30 000 retweets sur une même thématique est nécessaire.

En table 9.3, les valeurs de nombre d'abonnés et de popularité des deux comptes les plus influents (*jack* et *realDonaldTrump*) sont marquées d'une étoile, car nous ne disposons d'aucun tweet dont ils sont les auteurs : les nombres d'abonnés sont donc ceux constatés en juillet 2018, en forte augmentation depuis la période de collecte du corpus. L'influenceur $n^o$ 1 est donc *@jack* : Jack Dorsey, le fondateur de Twitter. Souvent mentionné, il est cependant (relativement) peu retweeté et n'a qu'un faible score d'expertise. Parmi notre Top10 se trouvent Donald Trump et son vice-président Mike Pence, trois sites d'information les soutenant : *BreibartNews*, *dcexaminer* et *DRUDGE_REPORT*, et des personnalités publiques (*benshapiro*, *iownjd*, *XplodingUnicorn*, *Sam___Hurley*).

Notons qu'à popularité presqu'égale, *iownjd* et *DRUDGE_REPORT* ont des expertises très différentes. Le premier produit des messages sur les modes de vie, les ressentis, la psychologie qui sont réémis mais ne se distinguent pas par leur originalité ; le second produit des contenus qui alimentent un débat politique intense, sur l'aile droite du parti républicain.

Le classement des comptes selon leur influence est certes basé sur les retweets et mentions ; cependant il est plus subtil qu'une simple somme. La table 9.4 liste les comptes les plus mentionnés dans

TABLEAU 9.3 – Top10 des comptes selon l'*influence*

| Nom | Abonnés | Popularité | Expertise |
|---|---|---|---|
| jack | 3 993 610* | 0.83* | 0.07 |
| realDonaldTrump | 53 184 614* | 0.97* | 0.76 |
| benshapiro | 384 447 | 0.7 | 0.88 |
| mike_pence | 886 620 | 0.74 | 0.92 |
| iownjd | 1 074 585 | 0.75 | 0.66 |
| XplodingUnicorn | 696 319 | 0.73 | 0.88 |
| BreitbartNews | 532 465 | 0.72 | 0.97 |
| DRUDGE_REPORT | 1 108 855 | 0.76 | 0.91 |
| dcexaminer | 81 544 | 0.61 | 0.82 |
| Sam___Hurley | 158 605 | 0.65 | 0.73 |

le jeu de données *KevRandTweets*. Dans ce tableau, la colonne « Rang » indique la position dans le classement des influenceurs par le *PageRank* sur le graphe des mentions (la position 0 correspondant à *jack* : réciproquement, un chiffre élevé correspondant à une influence faible). La colonne « Tweets » indique le nombre de tweets émis par ce compte, et présents dans le jeu de données. Il ne correspond donc pas directement à l'activité réelle du compte dans la période d'analyse : par exemple nous ne disposons d'aucun tweet de Donald Trump sur cette période (en conséquence, comme en table 9.3, le nombre d'abonnés de D. Trump indiqué ici ne correspond pas à sa valeur lors de l'émission des tweets, mais à la valeur constatée lors de la rédaction, juin 2018 ; cette situation est signalée par une étoile *). Dans ce corpus, *mike_pence* est certes moins mentionné que *RightWingAngel*, mais est nettement mieux reconnu en tant que référence par notre score d'influence.

TABLEAU 9.4 – Top10 des influenceurs selon le nombre de mentions

| Nom | Mentions | Rang | Tweets | Expertise | Abonnés | Popularité |
|---|---|---|---|---|---|---|
| RightWingAngel | 408 063 | 25 | 2 229 | 0.95 | 83 997 | 0.62 |
| realDonaldTrump | 302 215 | 1 | 0 | 0.76 | 53 184 614* | 0.97* |
| BreitbartNews | 253 200 | 6 | 1 195 | 0.97 | 532 465 | 0.72 |
| DRUDGE_REPORT | 249 771 | 7 | 1 647 | 0.91 | 1 108 855 | 0.76 |
| LindaSuhler | 174 880 | 14 | 4 814 | 0.88 | 273 981 | 0.68 |
| benshapiro | 160 649 | 2 | 2 367 | 0.88 | 384 447 | 0.70 |
| mike_pence | 160 576 | 3 | 68 | 0.92 | 886 620 | 0.74 |
| dcexaminer | 130 613 | 8 | 6 343 | 0.82 | 81 544 | 0.61 |
| CarmineZozzora | 110 849 | 42 | 8 390 | 0.8 | 172 182 | 0.65 |
| Lrihendry | 105 091 | 26 | 2 076 | 0.84 | 119 049 | 0.63 |

La nuance entre comptes actifs, mentionnés ou influents est nécessaire : pour comprendre ce que contient un corpus, il faut identifier à la fois les comptes qui sont souvent mentionnés, ainsi que ceux qui agissent le plus. Les indicateurs bruts (nombres d'abonnés, de tweets émis, de mentions faites) fournissent un premier élément de réponse.

Pour compléter ces indicateurs, nous avons introduit des scores : l'**influence** fournit un classement des comptes, selon l'importance de la position (dans le graphe des interactions) des utilisateurs qui les mentionnent. La **popularité** projette entre 0 et 1 les nombres d'abonnés finalement abstraits, facilitant la comparaison entre les comptes. Enfin, l'**expertise** répartit les retweets par thématique, ce qui récompense les comptes qui privilégient la communication ciblée.

L'utilisateur du système SARTN est ainsi doté d'outils d'exploration des acteurs-clés du jeu de données considéré, répondant au critère 2. Nous pensons qu'il existe plusieurs moyens d'être influent, ce qui implique des mesures différentes ; nos scores prennent en compte ces subtilités et différencient

les comptes populaires et les comptes fréquemment retweetés sur un domaine spécifique.

Nous étudions les comportements des comptes dans cette section, en exploitant la technique de construction de profils-types non-supervisés.

#### 9.2.4.2 Profilage : traitement initial sur les données

À partir du modèle de données défini en chapitre 6, qui répartit les caractéristiques mesurées en 5 aspects (Biographie, Style, Média, Interaction, Temporel), nous proposons une méthode de clustering (en l'occurrence, un k-moyennes) afin de rendre le modèle plus digeste et exploitable par l'utilisateur du système SARTN. Tout d'abord, certaines caractéristiques sont remplacées par leur logarithme, à cause de leur dispersion initiale. Dans une seconde phase, nous normalisons les données pour obtenir une distribution centrée en 0, d'écart-type 1. Enfin, une ACP [2] permet de réduire la dimension du problème. Appliquée sur le corpus *KevRandTweets*, ceci permet de passer de 27 caractéristiques à 5, en conservant 59% de la variance.

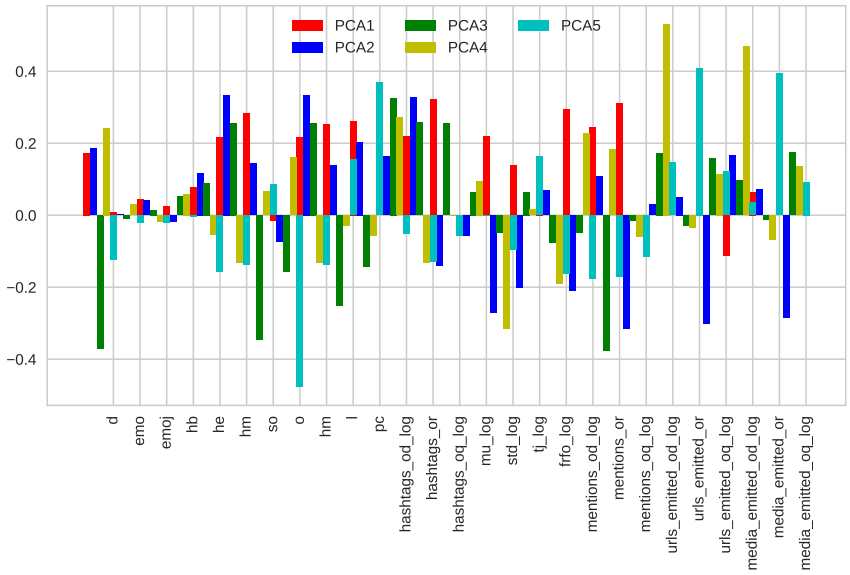



FIGURE 9.3 – Description des dimensions de l'ACP

L'ACP produit un nouvel espace le long de dimensions artificielles (les axes PCA1, PCA2... PCA5) grâce à des combinaisons linéaires des caractéristiques initiales. Pour mieux comprendre ce nouvel espace, la figure 9.3 expose le poids de chacune des caractéristiques initiales dans le calcul des nouvelles dimensions. Par exemple, *PCA1* (en rouge) dépend notamment de la quantité de hashtags, de mentions et de tweets par jour : de grandes valeurs de *PCA1* sont liées à la publication de tweets plus nombreux, contenant plus de hashtags et de mentions. *PCA2* (en bleu) est négativement liée avec la quantité totale de tags, mentions et objets sociaux. De son côté, *PCA3* (en vert sombre) est négativement liée aux scores de diversité, sociabilité et originalité. *PCA4* (en vert clair) est positivement liée avec la quantité de comptes différents mentionnés et avec le ratio quotidien d'émission d'URLs, et négativement liée à l'écart-type des temps entre deux publications. Finalement, *PCA5* (en cyan) est négativement liée à l'originalité (et donc, positivement liée à un comportement de retweet), et positivement liée au ratio de présence d'URLs et de média (images).

2. Analyse en Composantes Principales

#### 9.2.4.3 Calcul des profils-types / clusters

Sur la matrice des observations projetée dans l'espace de l'ACP, nous appliquons l'algorithme des k-moyennes pour attribuer un même label à des profils d'utilisateurs similaires, car proches. Le choix de la valeur de $k$ pourrait être arbitraire ; nous décidons d'utiliser la mesure de Calinski-Harabasz [Caliński and Harabasz, 1974], un ratio entre dispersions inter- et intra- clusters. Une valeur élevée signifie des clusters mieux formés, plus éloignés les uns des autres.

Après un calcul pour plusieurs valeurs de $k$, le meilleur choix est $k = 6$, qui maximise la valeur de la mesure de Calinski-Harabasz, comme l'illustre la figure 9.4.

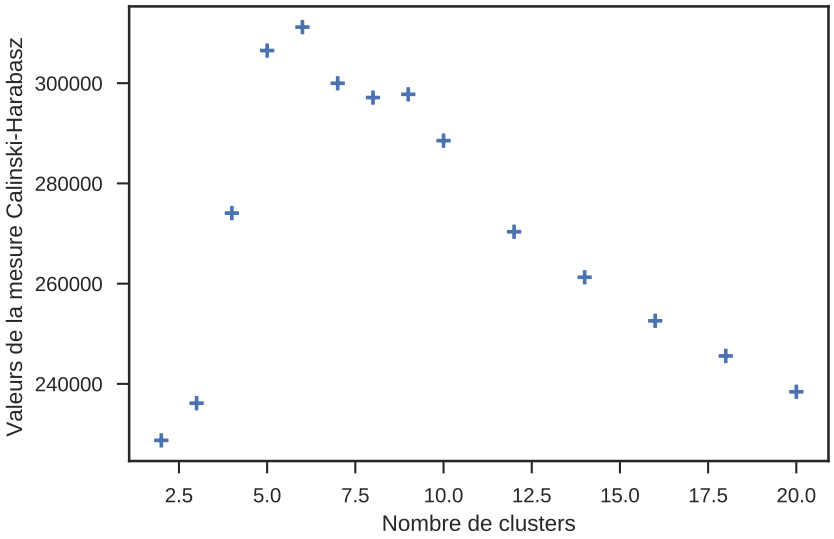



FIGURE 9.4 – Mesure de qualité des clusters / profils-types selon leur nombre

La figure 9.5 montre la répartition des clusters ou profils-types le long des deux premiers axes de l'ACP. L'interprétation se nourrit de la description des valeurs de l'ACP, via la figure 9.3, ou bien de l'analyse empirique des répartition des caractéristiques de chaque profil-type / cluster. La représentation en nuage de points est parfois trompeuse, car quelques clusters sont bien moins denses que les autres, tel celui en bas à droite (cluster 4 en bleu) : il est constitué d'utilisateurs très actifs, recourant en masse aux mentions et hashtags. Autour du point central (0,0) se trouvent les utilisateurs « normaux », de faible activité mais avec des nuances entre profils 0, 1 et 5. En haut de la figure sont placés les profils correspondant à un niveau normal à élevé d'activité, associé cependant à une faible utilisation des hashtags et mentions.

La répartition du nombre de comptes parmi les 6 clusters est plutôt équilibrée, compte tenu des écarts connus de niveaux d'activité déjà documentés [Kwak et al., 2010]. Le plus petit cluster contient 32 000 utilisateurs (cluster 4) ; le plus grand en compte près de 260 000 (cluster 5). Les clusters 0 et 1 dépassent les 130 000 comptes.

#### 9.2.4.4 Comparaison entre profil-type et score d'influence

Intuitivement, il y a un lien entre l'activité d'un compte et son influence : quelques mécaniques comportementales favorisent l'influence d'un compte. La figure 9.6 montre la répartition du score d'influence (comme défini en Équation 6.3), calculé à partir du graphe des interactions $G_I$ pour chacun des utilisateurs. Nous concentrons ici notre analyse sur les comptes sources, c'est-à-dire dont tous les messages ont été récupérés durant la période de collecte, et *influents*, c'est-à-dire dont le score $pr$ est supérieur à $5.0 \times 10^{-4}$. Ce seuillage correspond au top-5% des utilisateurs du jeu de données complet.

Par exemple, le cluster 4 (en bleu foncé sur la figure 9.6) est lié à une utilisation massive des mentions ; nous pouvons les qualifier « d'émetteurs ». Ceci a un impact sur leurs scores d'influence,

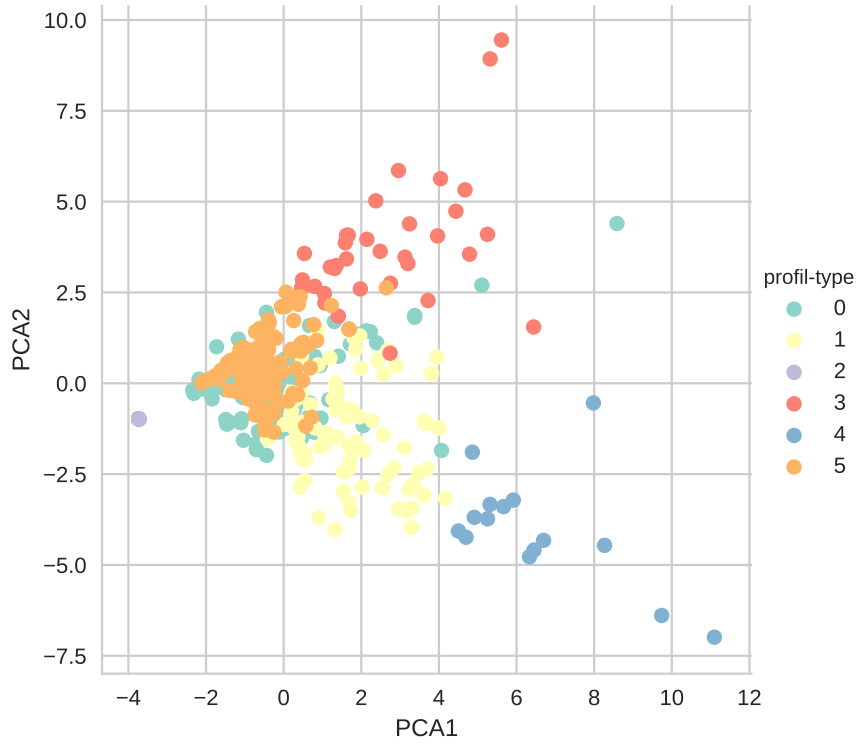



FIGURE 9.5 – Clusters / profils-types placés le long des deux premiers axes de l'ACP

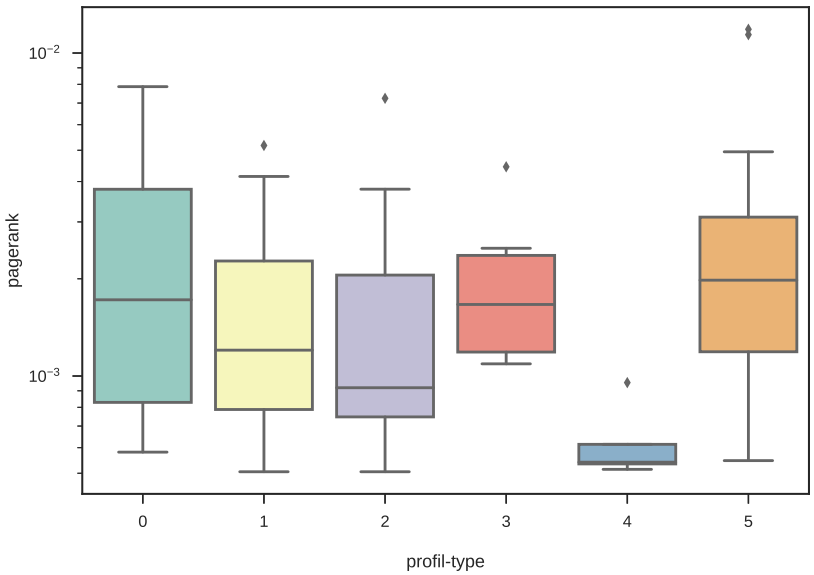



FIGURE 9.6 – Score d'influence VS Profil-type pour les comptes sources influents

car leur pageranks sont relativement faibles au regard des autres clusters : ils créent des liens vers d'autres comptes, mais ne bénéficient pas particulièrement de liens entrants. À l'opposé, le cluster 3 (en rouge) semble regrouper des comptes *influents* avec des scores élevés d'influence ; ces comptes sont liés à une activité régulière, non excessive.

Les profils-types 0 et 5 occupent le même espace en figure 9.5, mais le cluster 5 affiche des scores d'influence plus élevés. Finalement, la différence de comportement entre les clusters 1 et 2 ne résulte pas en une différence de score d'influence notable.

#### 9.2.4.5 Rôles-types et profils-types

Les **rôles** sont un niveau d'abstraction résumant les types de positions occupées par un utilisateur-nœud. Le système SARTN les calcule grâce à la méthode *RolX* [Henderson et al., 2012]. Pour chaque nœud, un ensemble de caractéristiques topologiques sont calculées, puis un algorithme de clustering non-supervisé attribue des étiquettes aux nœuds, les rôles-types, respectant le critère 3.

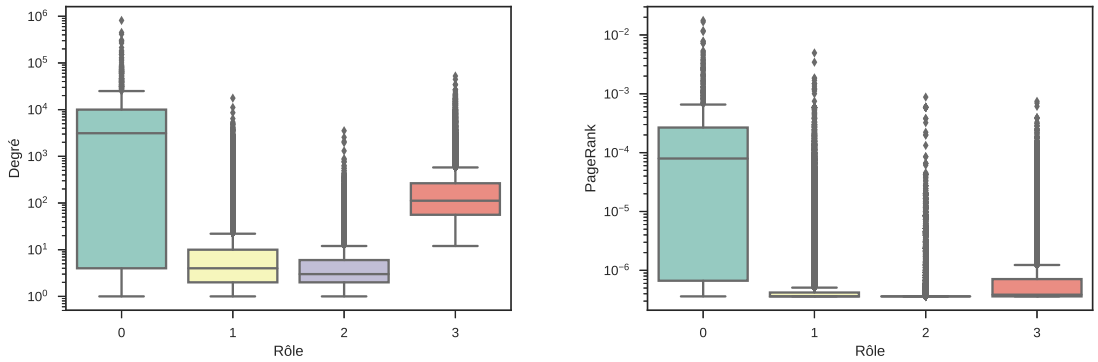



FIGURE 9.7 – Répartition des degrés et du PageRank selon le rôle-type

Sur le graphe des interactions, $G_I$, quatre rôles-types sont ainsi calculés et attribués aux nœuds-utilisateurs, résumant leur type de position : il revient à l'utilisateur de SARTN de qualifier les rôles-types à l'aide des caractéristiques topologiques explicatives, fournies elles aussi par *RolX*, pour étiqueter ces rôles-types en nœuds-feuilles, émetteurs, récepteurs, ponts, par exemple. La figure 9.7 accompagne cette démarche d'interprétation, en présentant la répartition du degré (à gauche) et du PageRank (à droite) selon le rôle-type des nœuds. Le rôle 0 correspond à des nœuds très centraux, échangeant beaucoup avec leur environnement ; de fort degré et PageRank, ils ne sont pas nombreux. Le rôle 1 est associé à un niveau plus faible de connexions, tant entrantes que sortantes. Le rôle 2 est attribué à des nœuds-feuilles, ou récepteurs purs : ils n'émettent pas d'interactions mais en reçoivent quelques-unes. Leur PageRank est généralement très faible. Enfin, le rôle 3 représente des nœuds surtout récepteurs, mais bien reliés à leur environnement (de degré élevé, ils sont mentionnés par plusieurs autres comptes).

La comparaison des répartitions des profils-types et des rôles-types ne montre pas de corrélation entre les deux. Cependant le nombre de catégories (4 rôles-types, mais 6 profils-types) ainsi que la présence de comptes très actifs en bordure de corpus (acteurs-clés dont les contenus n'ont pas été ciblés lors de la collecte) sont susceptibles d'avoir un impact sur cette corrélation.

Ces indicateurs d'influence, de comportement et de position sociale permettent de bien caractériser les comptes utilisateurs, et fournissent à l'utilisateur de SARTN, tous les éléments pour accomplir sa mission. Bien sûr, des tâches spécifiques peuvent rendre nécessaire l'approfondissement de ces indicateurs, pour établir un profil psychologique, émotionnel, ou encore pour mieux prendre en compte la dynamique des actions et l'évolution de la position sociale de l'individu. Ces informations sont encore exploitables pour éclaircir le fonctionnement des entités « communautés », point focal de la

section suivante.

### 9.2.5 Détection et caractérisation des communautés

Dans cette section, nous nous penchons sur les groupes détectés à partir du graphe des interactions, introduit en section 9.2.3. Ce graphe est noté $G_I$ et contient tous les liens d'interaction directe publique, c'est-à-dire les retweets, les mentions et les réponses.

#### 9.2.5.1 Répartition des profils-types au sein de groupes

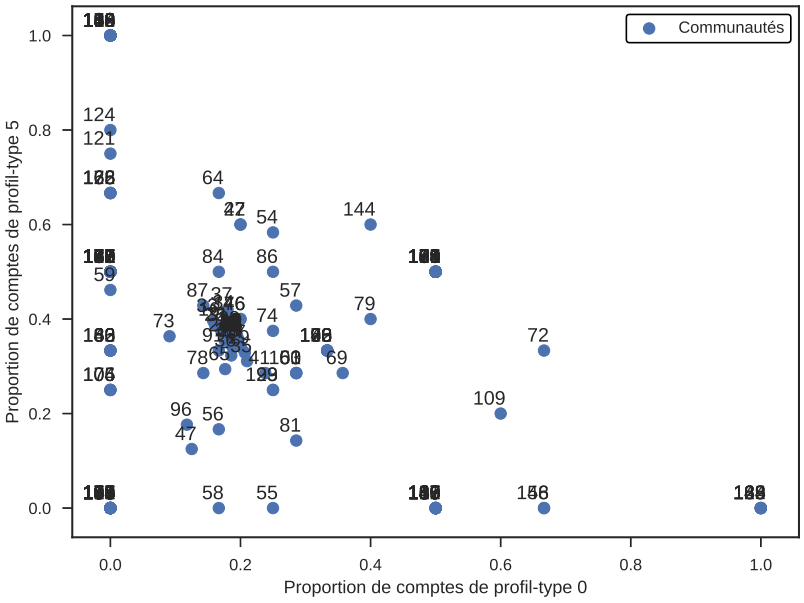



FIGURE 9.8 – Répartition des profils-types au sein des communautés détectées sur $G_I$

L'exploitation des profils-types permet de mieux caractériser des communautés détectées à partir d'un graphe. La figure 9.8 montre le pourcentage de comptes de profil-type 0 en abscisses, contre le profil-type 5 en ordonnées : il s'agit des deux profils-types les plus fréquents dans notre étude. Chaque point correspond à une communauté détectée. De nombreux groupes apparaissent autour des coordonnées $(0,19;0,35)$, et constituent *de facto* des groupes « normaux ». Cette position correspond aux proportions fréquentes de ces profils-types dans la population.

Cependant, d'autres points d'équilibre sont visibles entre ces deux types de profils : nous avons l'intuition qu'il existe des « types de groupes », basés sur les comportements de leurs membres, et correspondant à différents niveaux d'activité et d'engagement de la part des utilisateurs.

#### 9.2.5.2 Étude de cas : analyse de quelques groupes de comptes

Afin d'illustrer l'exploitation des mesures topologiques et thématiques, nous proposons de « zoomer » à l'intérieur de quelques groupes, nommés par des IDs : *group84*, *group69*, *group26* et *group3*. Ces groupes ne sont pas choisis au hasard : ils illustrent des situations couvrant un large morceau du domaine des mesures topologiques et thématiques. Les groupes 84 et 69 partagent une petite taille et un faible ρ, tandis que les groupes 84 et 26 ont un bon $\theta f.igf$ en commun. Les groupes 3 et 26 ont tous deux des ρ élevés, et un grand nombre de membres.

La table 9.5 expose leurs statistiques : par exemple, le *group84* est composé de 6 membres ; bien que seulement un tiers d'entre eux soient marqués comme actifs sur le même sujet, ce fait le distingue

TABLEAU 9.5 – Statistiques descriptives de 4 groupes

| | Mesures topologiques | | | | Mesures thématiques | | | |
|---|---|---|---|---|---|---|---|---|
| ID | taille | d | c | TPR | ξ | ρ | $\theta f.igf$ | $\xi_{sim}$ |
| 84 | 6 | 0.46 | 0.22 | 0.5 | 0.33 | $> 10^{-4}$ | 0.77 | 0.93 |
| 69 | 14 | 0.16 | 0.0 | 0.21 | 0.07 | $> 10^{-4}$ | 0.21 | 0.88 |
| 26 | 2 083 | $> 10^{-3}$ | 0.02 | 0.97 | 0.64 | 0.14 | 2.53 | 0.96 |
| 3 | 51 208 | $> 10^{-4}$ | 0.31 | 0.60 | 0.12 | 0.18 | 0.26 | 0.90 |

(bon score $\theta f.igf = 0,77$). Ce groupe est formé autour de retweets qui sont suivis de réponses émises par le compte retweeté.

La figure 9.9 montre une communauté, le **groupe 84**, ainsi que les nœuds et arcs l'entourant : plus petits, en noir, ce sont les portes vers l'extérieur de la communauté, c'est-à-dire vers le reste du réseau. La partie épaisse des arcs sont les pointes des flèches. La figure n'est qu'une vue très partielle du graphe complet. La spatialisation est calculée grâce à l'algorithme ForceAtlas2 : les positions (x,y) ne portent pas de signification particulière ; l'algorithme éloigne les nœuds les uns des autres, tandis que les arcs fonctionnent comme des ressorts, rapprochant leurs deux extrémités. Les couleurs sont issues de la détection non-supervisée de rôles des nœuds : comme précisé précédemment, les nœuds verts sont des comptes plutôt émetteurs de liens, en connexion avec leur environnement, tandis que les bleus ont des positions de feuilles, récepteurs d'un petit nombre de liens, et n'en émettant pas ou peu.

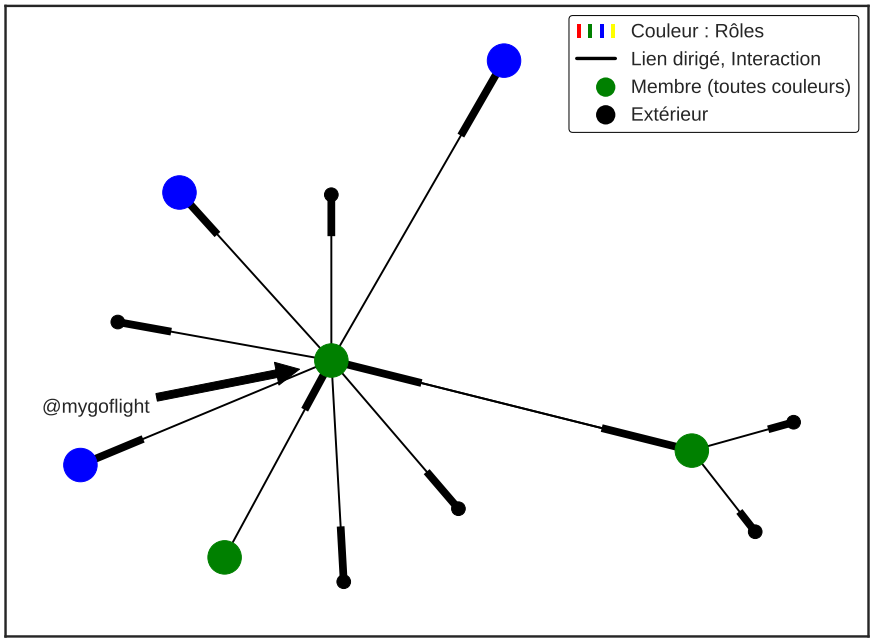



FIGURE 9.9 – Vue du groupe 84

L'utilisateur central, en vert, est nommé @mygoflight. Il s'agit d'un vendeur de produits liés au pilotage, à l'aviation. Pour réaliser sa vocation commerciale, il émet des contenus promotionnels, qui sont à leur tour retweetés par des clients passés ou futurs. La visualisation des rôles permet d'écarter les trois nœuds bleus, qui ne participent pas à la vie du groupe.

Les valeurs élevées des scores topologiques (densité, TPR) du *groupe84* sont expliquées par les retours faits par l'utilisateur central vers ses retweeteurs : la plupart du temps, il s'agit d'un message de remerciement pour le retweet. La conductance moyenne résulte de la connexion très forte de l'un des membres avec un blog politique. Le $\theta f.igf$ élevé s'explique par la spécificité de la thématique (des références à des produits précis), en nombre suffisant grâce aux retweets de contenus promotionnels. En guise de conclusion, le groupe 84 est focalisé sur une thématique, montrant une quantité raisonnable d'interaction interne.

Bien que plus grand, le **groupe 69**, illustré en 9.10, est lui aussi organisé autour d'un compte commercial ; sa thématique principale (bien que minoritaire) concerne le hashtag *#AWholeLotOfChristmas* (« un gros morceau de Noël », une initiative de collecte de fonds pour une association, via des décorations et illuminations de maisons, état de Géorgie, USA). La communauté détectée existe en grande partie par l'action continue menée par son nœud central, *@designyourtie*, qui vend notamment des cravates personnalisées : promotion de produits, messages de remerciement aux consommateurs satisfaits, et vérification de la satisfaction client, couvrant des thématiques humainement proches, mais dont les messages ne partagent aucun mot, donc différentes au sens de LDA. La conductance, à zéro, est due à l'absence totale d'arcs sortants. Les nœuds externes visibles en noir en figure 9.10, n'accomplissent qu'un faible nombre d'actions envers des membres de la communautés.

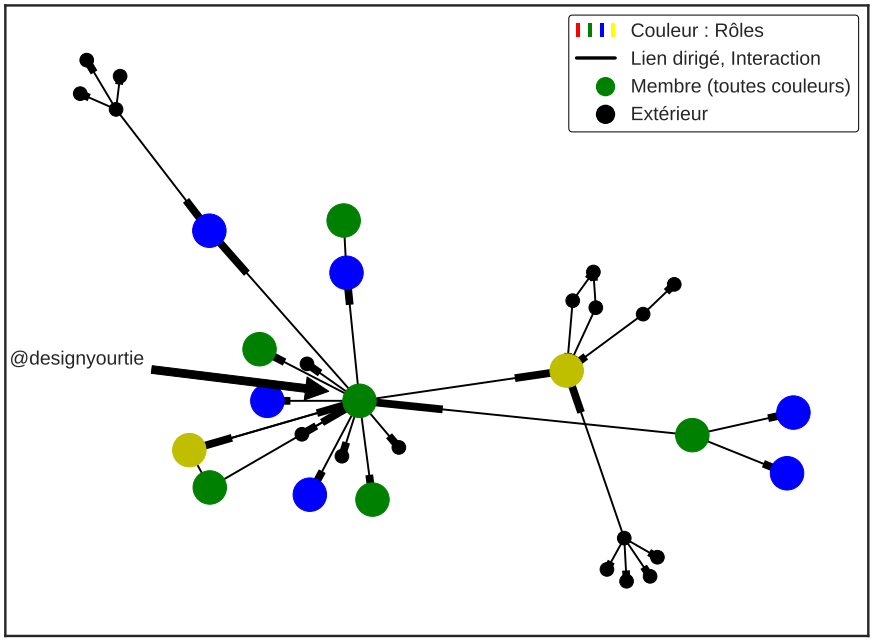



FIGURE 9.10 – Vue du groupe 69

Au centre principal (*@designyourtie*) s'ajoutent des centres secondaires, tel le nœud jaune situé sur la droite. Ce rôle correspond à une position à la fois source et réceptrice d'interactions, ici positionnée à l'interface vers d'autres communautés. Les chapelets de nœuds extérieurs qui l'entourent, colorés en noir, résultent de messages mentionnant plusieurs comptes, typiques des tentatives d'accroches supposées favoriser l'ajout d'abonnés : en un tweet apparaissent des liens dirigés vers cinq nœuds.

Pour continuer avec l'une des grandes communautés détectées, le **groupe 26**, illustré en figure 9.11, est constitué des interactions émanant d'un ensemble de tweets politisés, qu'ils soient pro- ou anti- Donald Trump. Les principaux membres, en termes de nombre d'abonnés, incluent des producteurs de contenus apolitiques (@*iamoppose*, @*AliceMartin8*, @*Oceaanfietser*, qui émettent des contenus *feelgood* -vie et bien-être, ainsi que des *memes* : images à textes ou GIF), et des références politiques majeures telles que @*wikileaks* ou encore @*netanyahu*, le premier ministre israélien.

Le principal compte de ce grand groupe n'est pas pro-Trump (à l'opposé de la majorité des membres de ce groupe, et du corpus) : @*annemariayritys* revendique soutenir le parti démocrate, et déclenche à la fois des messages de soutiens et d'opposants. La faible conductance de ce groupe provient du petit volume d'arcs internes, complètement dépassé par la quantité d'arcs externes : la plupart des comptes dialoguent avec des membres d'autres communautés.

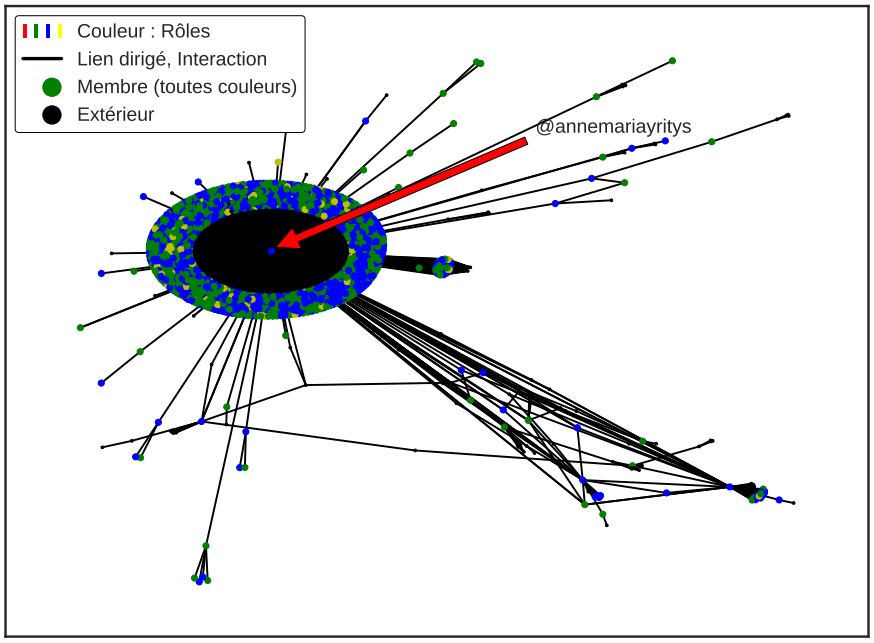



FIGURE 9.11 – Vue du groupe 26

Ici encore, les nœuds sont colorés selon leur rôle. Cependant, le groupe-graphe contient trop de nœuds (un total de 2 237), et sa visualisation n'est pas pertinente. Pour analyser un graphe, il faut pouvoir interagir avec la vue pour l'explorer activement, en filtrant et en travaillant la spatialisation. Les outils de sélection des comptes à visualiser, de filtrage selon le degré des nœuds ou de poids des arcs sont à exploiter ici.

Finalement, nous analysons le **groupe 3**, qui est très gros, avec plus de 50 000 membres. Il est dispersé thématiquement, et structurellement multi-centré : les nœuds de haut degré incluent Mike Pence (Vice-Président des États-Unis), Donald Trump, son fils Donald Trump Junior, les anciens candidats républicains à la primaire Ben Carson et Ted Cruz, ainsi que le magazine Time, qui servent de références au reste de la communauté. Le cœur du groupe est composé d'émetteurs très actifs, qui mentionnent beaucoup ces références dans leurs tweets. Le groupe est trop gros pour en proposer une visualisation exploitable, comme fait pour les trois groupes précédents.

Ces groupes sont visibles en figure 9.14, indiqués par des croix rouges. Ils sont placés aux quatre

coins du graphique, afin de donner des exemples très différents des communautés détectées. Ces quatre exemples illustrent la performance de SARTN vis-à-vis des critères 5 et 6, portant respectivement sur la détection des communautés, et sur le couplage des mesures topologiques et thématiques.

#### 9.2.5.3 Mesures topologiques de la force de l'interaction

Dans chacune des figures suivantes, la taille des points est logarithmiquement proportionnelle au nombre de nœuds dans une communauté. La figure 9.12 montre la densité interne des groupes comparée à leur ratio de participations à des triangles (TPR). Les densités sont habituellement basses pour les grands groupes, car le nombre d'arcs possibles est quadratique en comparaison du nombre de nœuds, c'est pourquoi l'échelle logarithmique est retenue ici. Le long de l'axe TPR, les groupes sont répartis entre faiblement connectés (principalement, des petits comptes qui ont retweeté une fois une référence : un blog ou une personnalité politique), et fortement connectés : les membres du groupe ont tous quelques amis au sein du groupe. En guise de conclusion, la densité est trop liée à la taille du groupe, tandis que le TPR illustre de vraies différences entre groupes de même taille.

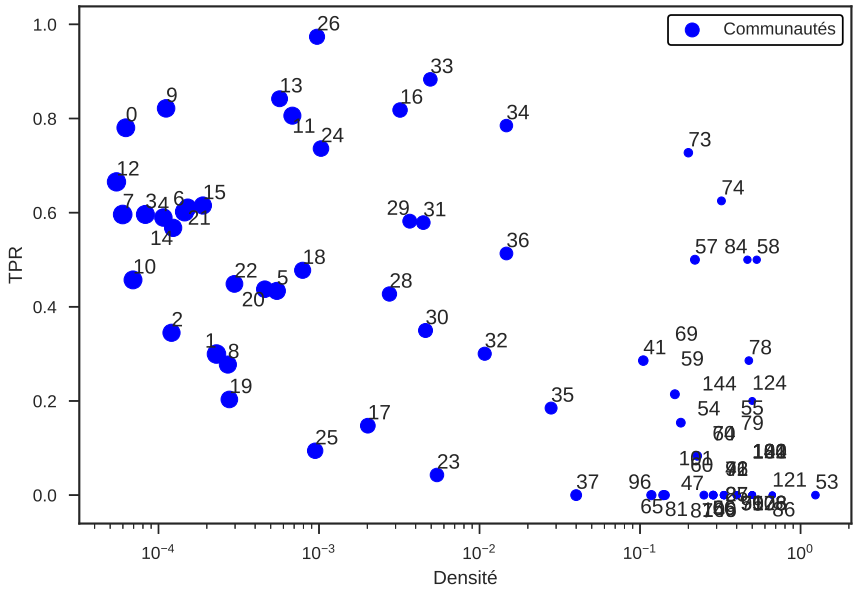



FIGURE 9.12 – Comparaison des densités et ratio de participation à des triangles (TPR)

La cohésion interne est comparée avec la quantité de liens vers l'extérieur en figure 9.13. Sur l'axe des ordonnées, la conductance mesure la quantité relative de liens vers l'extérieur, à la frontière du groupe. Une forte conductance suggère qu'un groupe est fortement connecté avec le reste du réseau. Le quadrant inférieur droit contient donc les groupes montrant une forte interaction interne, avec une faible conductance : en quelque sorte, ils sont isolés de la circulation d'information du cœur du réseau. Le comportement opposé est observé dans le quadrant supérieur gauche : un faible TPR et une forte conductance implique des groupes avec très peu de liens internes.

#### 9.2.5.4 Mesures thématiques de pertinence du groupe

La figure 9.14 illustre deux mesures que nous avons introduites : l'*expertise* $\xi$ en abscisse représente la proportion de membres d'un groupe, actifs sur la même thématique ; la *représentativité* $\rho$ en ordonnées représente la proportion d'utilisateurs (dans le réseau) intéressés par la même thématique, et présents dans le groupe. Ainsi, les grands groupes contiennent la plupart des comptes d'une thé-

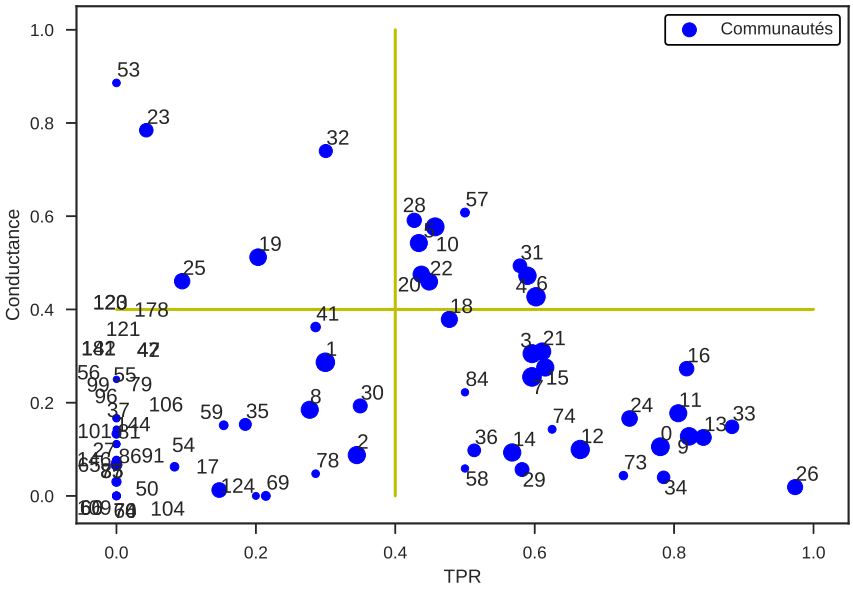



FIGURE 9.13 – Comparaison du TPR et de la conductance

matique (ρ élevé), mais sont aussi plus gros, plus inclusifs : leurs faibles ξ illustrent leur dispersion thématique.

Les groupes sont présents sur un intervalle allant de valeurs de ξ très faibles sur la gauche, correspondant à des communautés thématiquement dispersées, jusqu'à des groupes plus cohésifs, focalisés sur un thème, sur la droite du graphique. Une tendance globale est visible, même si elle tolère des exceptions : le point le plus visible parmi les groupes les plus aberrants, dans le quadrant supérieur droit de la figure 9.14, a été détaillé : il s'agit du groupe 26, thématiquement et structurellement dispersé. De fait, la plupart des grandes communautés sont multi-centrées et thématiquement dispersées.

La mesure ξ n'est pas l'unique manière d'évaluer la pertinence d'un groupe sur une thématique. $\theta f.igf$ relève aussi ce défi, inspirée du $tf.idf$ bien connu. La figure 9.15 illustre ce score sur l'axe des abscisses, contre ρ en ordonnées. Le graphique est globalement similaire à la figure 9.14, avec plus de dispersion le long de l'axe des abscisses : les groupes peuvent avoir un faible score ξ, mais sur une thématique bien particulière, peu courante dans le corpus. Cela suffit parfois à distinguer l'importance du groupe sur une telle thématique.

#### 9.2.5.5 Vue de la cohésion thématique et d'interaction

Puisque nous cherchons des groupes à la fois basés sur l'interaction, et focalisés sur une thématique, la meilleure représentation rassemble le ratio de participation à des triangles TPR, pour évaluer la force d'interaction interne, et $\theta f.igf$, pour la cohésion thématique.

La figure 9.16 compare les communautés détectées. Les lignes jaunes suggèrent une qualité minimale pour des groupes pertinents, cohésifs et thématiques, dans le quadrant supérieur droit. La méthode de partitionnement de graphe, basée uniquement sur les liens d'interaction, semble ne repérer que quelques groupes bien soudés (avec des TPR élevés), mais ceux-ci sont souvent dispersés thématiquement.

Les communautés isolées, sur le côté droit, incluent des petits groupes qui sont les seuls à participer sur leurs thématiques ; il s'y trouve aussi un grand groupe, déjà passé en revue (le groupe 26). Dans l'ensemble, les groupes sont dispersés tant sur l'axe des abscisses que sur celui des ordonnées. Les mesures que nous avons introduites sont focalisées sur la dimension thématique, tandis que les mesures de l'état de l'art sont habituellement topologiques. Nous pensons qu'il y a complémentarité,

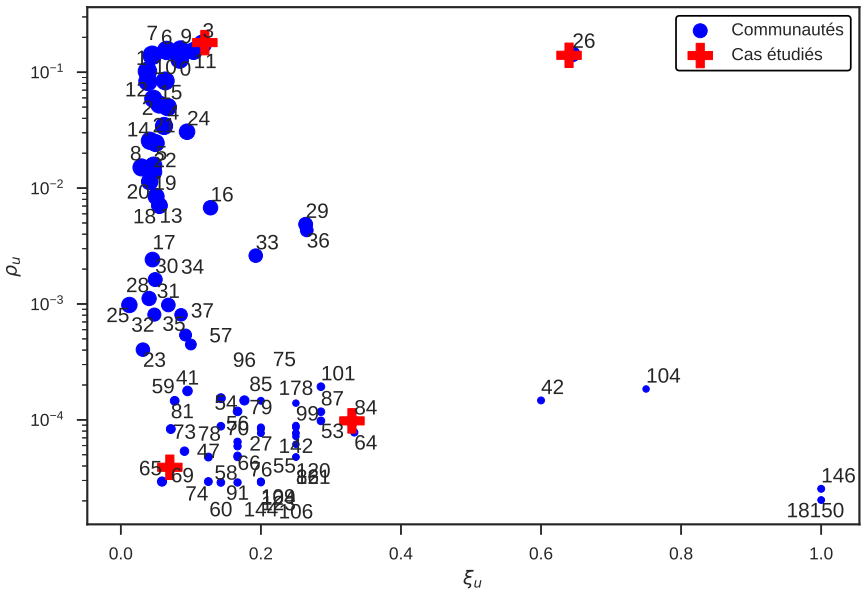



FIGURE 9.14 – Comparaison des scores $\xi_u$ et $\rho_u$

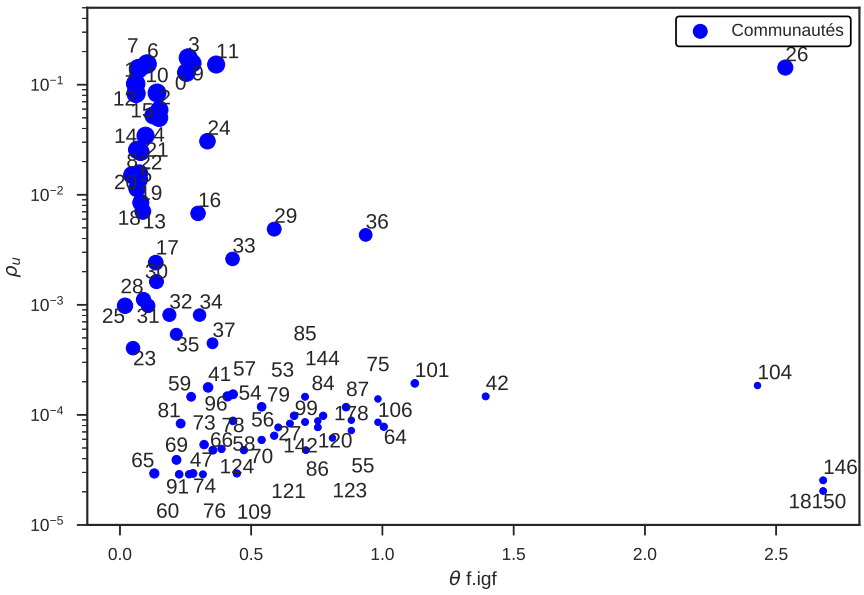



FIGURE 9.15 – Comparaison des scores $\theta f.igf$ et $\rho_u$

et que leur combinaison produit une vue du réseau plus informative.

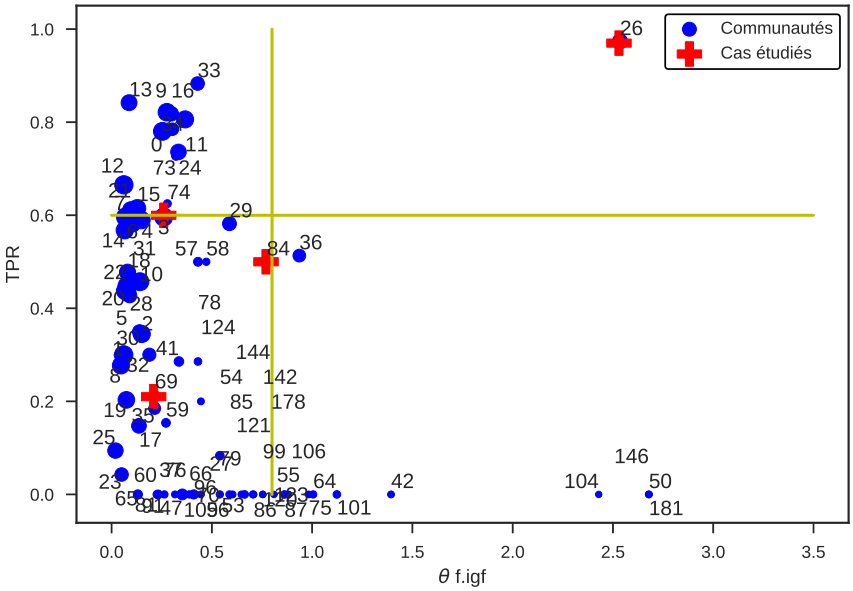



FIGURE 9.16 – Comparaison des scores $\theta f.igf$ et TPR

TABLEAU 9.6 – Description complète des scores atteints par les groupes sur *KevRandTweets*

| | Mesures thématiques | | | | Mesures topologiques | | | |
|---|---|---|---|---|---|---|---|---|
| | ξ | ρ | $\theta f.igf$ | $\xi_{sim}$ | d | c | TPR | taille |
| moyenne | 0.189 | 0.021 | 0.539 | 0.907 | 0.221 | 0.181 | 0.280 | 8 959 |
| min | 0.012 | $10^{-5}$ | 0.019 | 0.871 | $10^{-5}$ | 0.0 | 0.0 | 4 |
| 25% | 0.062 | $10^{-4}$ | 0.139 | 0.891 | $10^{-4}$ | 0.003 | 0.0 | 5 |
| 50% | 0.135 | $10^{-4}$ | 0.344 | 0.901 | 0.154 | 0.111 | 0.192 | 14 |
| 75% | 0.25 | 0.011 | 0.708 | 0.913 | 0.40 | 0.268 | 0.554 | 4 137 |
| max | 1.0 | 0.175 | 2.679 | 1.00 | 1.00 | 0.885 | 0.973 | $10^5$ |

La table 9.6 présente la répartition des caractéristiques thématiques et topologiques. Les valeurs moyennes et médianes peuvent être exploitées pour comparer différents algorithmes de détection de communautés. Comme les groupes sont parfois très petits, ρ atteint quelquefois des valeurs très faibles ; à l'opposé, pour des groupes très grands, c'est la densité qui est très faible ($10^{-5}$ environ).

#### 9.2.5.6 Similarité de thématiques dans la cohésion des groupes

Jusqu'à présent, les thématiques étaient considérées comme des étiquettes : un utilisateur est soit présent sur un thème, soit absent. Cependant, certaines thématiques sont similaires entre elles, se recouvrent parfois. Pour pallier ce problème, la mesure $\xi_{sim}$ prend en compte une similarité entre thématiques dans le calcul de la cohésion.

La figure 9.17 compare l'expertise classique ξ en abscisses, contre sa version utilisant la similarité thématique, $\xi_{sim}$, en ordonnées. La figure ne montre que les groupes de faible ξ, pour mieux accompagner notre commentaire. Comme attendu, les deux scores sont corrélés, mais présentent quelques caractéristiques intéressantes : dans le coin inférieur gauche, signalé par une ligne rouge en pointillés, des groupes de même $\xi = 0,10$ ont leur $\xi_{sim}$ dans un intervalle entre $0,87$ et $0,91$ : cette dernière valeur correspond d'habitude à des groupes de $\xi = 0,2$ : deux fois leur score initial. Cela signifie que

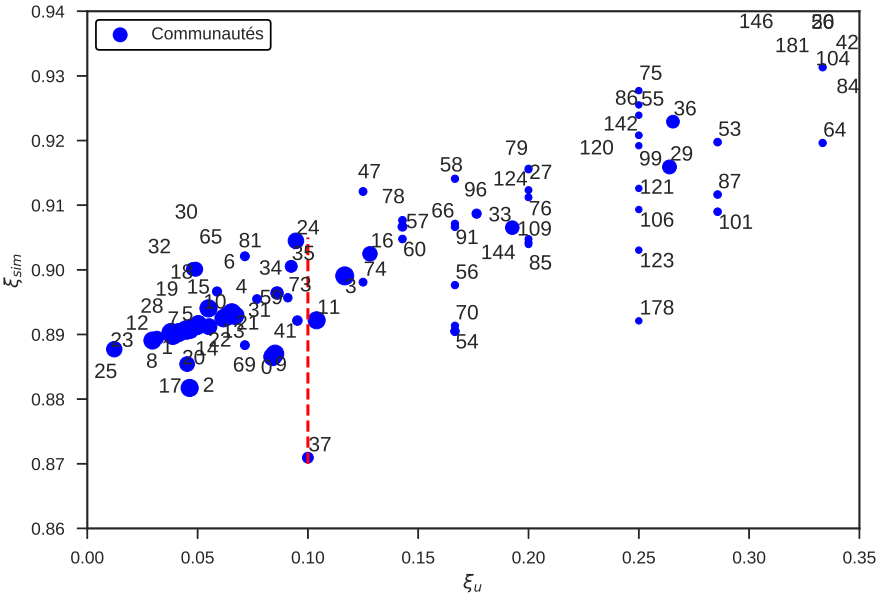



FIGURE 9.17 – Comparaison des scores $\xi_u$ et $\xi_{sim}$ : zoom sur les faibles valeurs de $\xi_u$

dans ce groupe placé en $(\xi = 0,1; \xi_{sim} = 0,91)$, les messages font partie de thématiques proches, et le groupe présente une cohésion thématique plus forte qu'à première vue, si seul $\xi$ est considéré.

Enfin, la figure 9.18 compare $\xi_{sim}$ et $\rho_u$. Ce graphique ressemble aux figures 9.14 et 9.15 ; cependant il disperse mieux les groupes les plus petits le long de l'axe des abscisses, $\xi_{sim}$, distinguant les discussions autour de thèmes proches d'une part, des ensembles de comptes utilisateurs sans intérêts partagés d'autre part.

Dans nos expériences, nous avons fixé la méthode de détection de thématiques, en optant pour LDA. L'existence d'une mesure de similarité n'est pas une évidence pour toutes les détections de thématiques ; cependant LDA est associée à une mesure de similarité entre textes et entre thématiques, qui facilite l'élaboration de la mesure $\xi_{sim}$. Ici, $\xi_{sim}$ estime la cohésion des communautés avec plus de subtilité que $\xi_u$, prenant en compte (et pondérant) toutes les thématiques plutôt qu'en se focalisant sur la principale.

#### 9.2.5.7 Cohésion en termes de comptes ou de textes publiés

Introduites en chapitre 7, les mesures $\xi$ et $\rho$ se déclinent de deux manières : en considérant les comptes utilisateurs, ou les textes publiés au sein du groupe. Une communauté est cohésive lorsque ses membres s'expriment en majorité sur une même thématique, ou bien lorsque la majorité des contenus échangés traitent du même sujet ?

Ainsi, $\xi_u$ et $\rho_u$ s'intéressent à la quantité de comptes *utilisateurs* focalisés sur une thématique au sein de la communauté, tandis que $\xi_t$ et $\rho_t$ dénombrent la quantité de *textes* émis, traitant de la thématique phare. Ces indicateurs reflètent une même réalité : la cohésion thématique des groupes. Dans cette section, les deux approches sont comparées pour vérifier leur similarité et leurs différences.

La figure 9.19 compare les *expertises*, mesurées en dénombrant les utilisateurs, avec $\xi_u$ (en abscisses), ou les textes avec $\xi_t$ (en ordonnées). La majorité des groupes se trouvent dans la zone des faibles scores, en bas à gauche. Si la tendance globale montre une corrélation entre les deux mesures, il se trouve toutefois des communautés de $\xi_u = 0.2$, c'est-à-dire modérément cohésive, qui affichent pourtant un $\xi_t > 0.6$ : quelques membres se chargent d'émettre une grande quantité de messages d'une même thématique. Nous pensons qu'il sera possible d'identifier ce type de communautés grâce

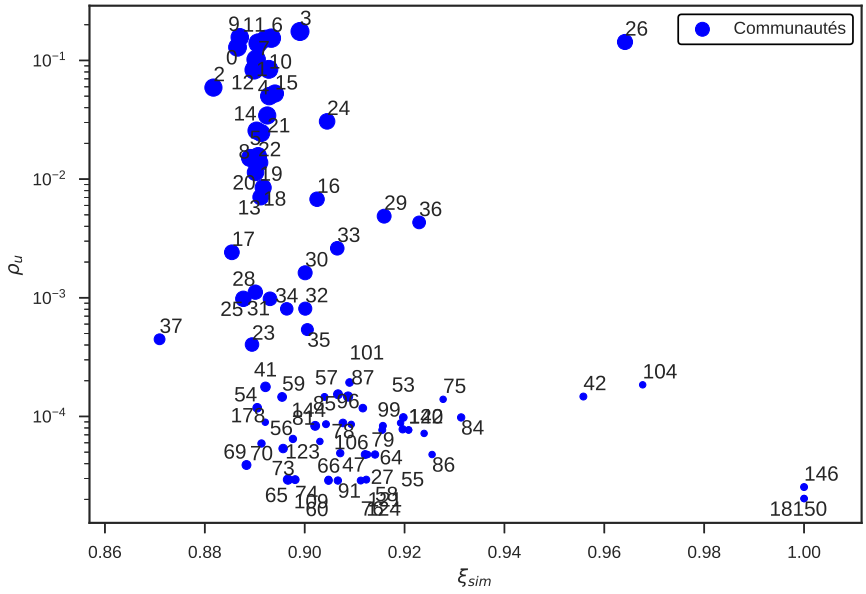



FIGURE 9.18 – Comparaison des scores $\xi_{sim}$ et $\rho_u$

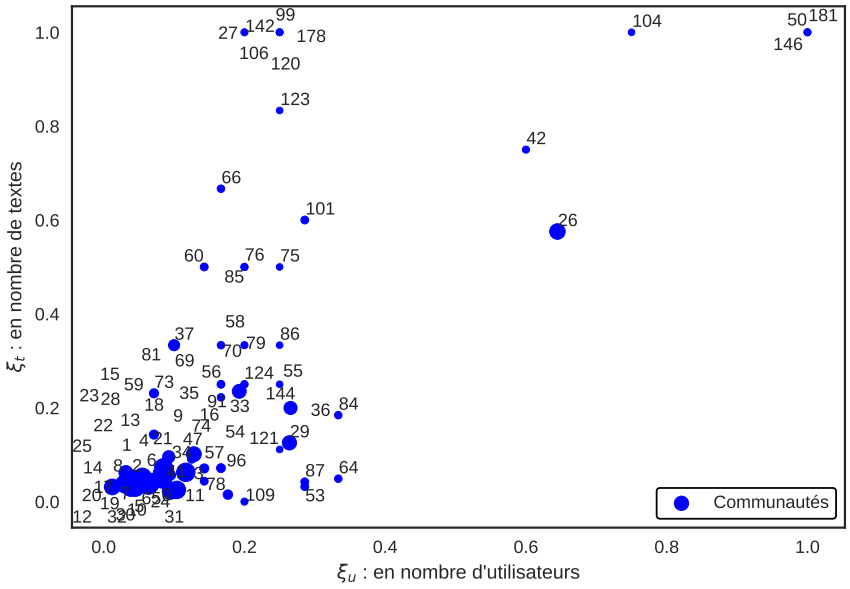



FIGURE 9.19 – Comparaison des scores $\xi_u$ et $\xi_t$

à l'étude de ces valeurs.

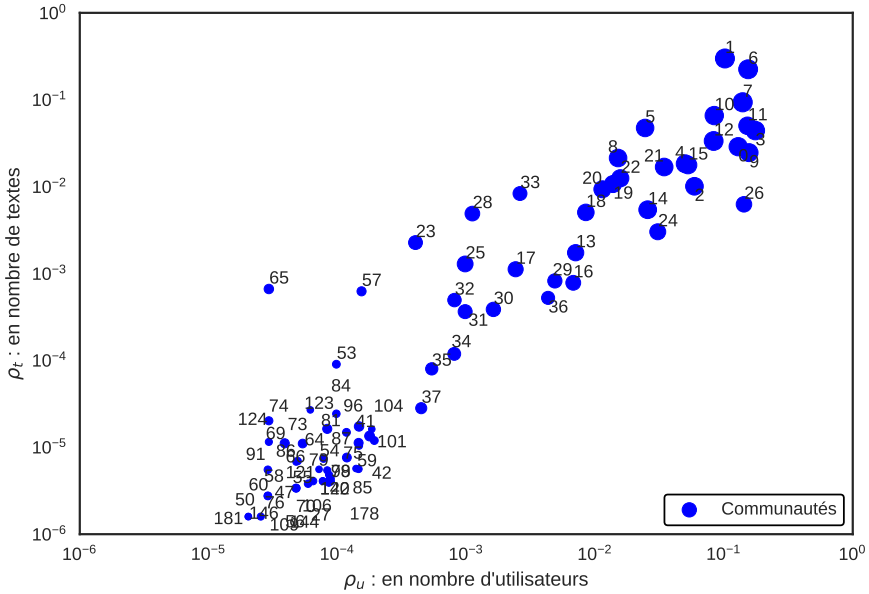



FIGURE 9.20 – Comparaison des scores $\rho_u$ et $\rho_t$

La figure 9.20 s'intéresse aux *représentativités*, mesurées en dénombrant les utilisateurs, avec $\rho_u$ (en abscisses) ou les textes, avec $\rho_t$ (en ordonnées). Comme d'habitude avec $\rho$, l'échelle des axes est logarithmique. Les deux mesures apparaissent clairement liées, bien qu'autorisant de nombreux écarts : à une valeur de $\rho_t = 10^{-3}$ correspond un intervalle de $\rho_u \in [10^{-5}; 10^{-2}]$. À ces écarts correspondent des situations différentes : certains comptes, souvent automatisés, publient de nombreuses fois des messages similaires voire identique (et donc de même thématique), ce qui augmente leur score $\rho_t$. D'autres groupes sont constitués autour de l'action de *retweet* d'un message spécifique : les membres appartiennent au groupe par leur retweet, qui est bien sûr de la même thématique. Ces membres qui n'ont publié qu'un unique message contribuent fortement à $\xi_u$, mais faiblement à $\xi_t$.

Enfin, la figure 9.21 affiche les groupes avec $\xi_t$ en abscisses et $\rho_t$ en ordonnées. Le nuage de points est similaire à celui de la figure 9.14 : les grandes communautés ont souvent des scores de cohésion faibles ($\xi_t < 0.20$), mais des $\rho_t$ élevés. Les groupes plus petits sont parfois plus cohésifs, mais leur taille limite leurs scores de représentativité.

Pour conclure, ces différentes méthodes de calcul de $\xi$ et $\rho$ mesurent bien le même phénomène, qu'elles soient basées sur les nombres de membres ou bien sur les quantités de textes émis. Elles permettent de quantifier la dispersion thématique des communautés, et leur participation sur des thèmes. La comparaison des mesures sur textes et sur utilisateurs constitue un prisme d'analyse des constituants du groupe : ainsi sont différenciés les présences de gros émetteurs, et de petits répéteurs. Cependant il est plus direct de qualifier ces comportements à l'aide des profils de comportements établis en chapitre 6.

### 9.2.6 Synthèse de l'étude de *KevRandTweets*

Dans cette section, nous avons présenté un corpus de près de 10 000 000 de tweets (très exactement 9 671 711), dont les textes ont été analysés pour en extraire automatiquement les thématiques et les sentiments. Bien que ces traitements contiennent des erreurs de classification, ils permettent toutefois de présenter une vue statistique des popularités et opinions des comptes vis-à-vis de mots-clés

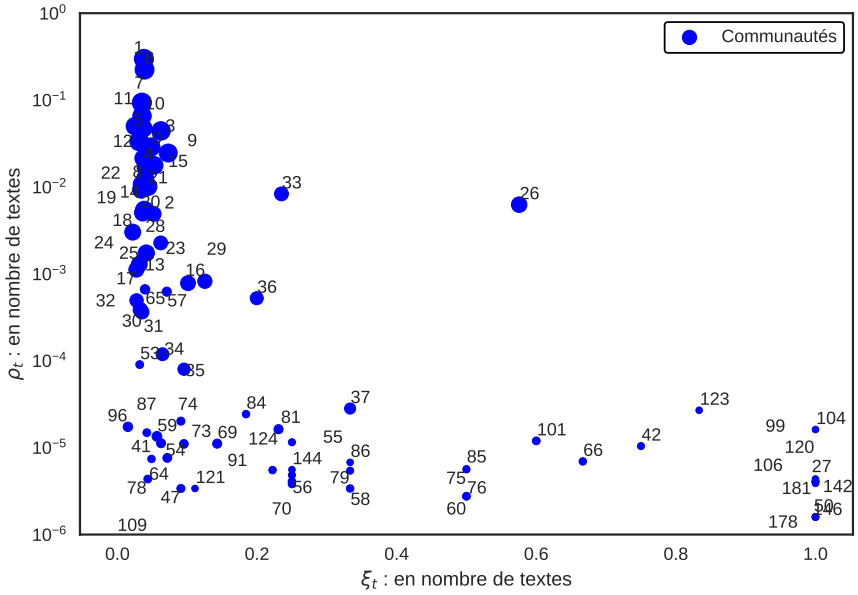



FIGURE 9.21 – Comparaison des scores $\xi_t$ et $\rho_t$

ou d'entités.

Les différents scores d'influence introduits identifient les acteurs-clés du corpus, en termes de popularité, d'influence et d'expertise sur une thématique. L'étude des comptes utilisateurs passe aussi par leur comportement (profil-type), et l'analyse de leur positionnement social (rôle-type). Ces informations, ainsi que les mesures de cohésion thématique ($\xi$, $\rho$, $\theta f.igf$) et structurelle, alimentent l'étude des communautés détectées dans le graphe des interactions.

Twitter est certes le plus emblématique des médias sociaux, par son ouverture et son instantanéité ; cependant d'autres plate-formes similaires existent, parmi lesquelles Galaxy2 (à une échelle bien différente). La seconde étude de cas va exploiter les mêmes outils sur ce second réseau social, montrant que ces outils ne sont pas spécifiques à une unique plate-forme : ils sont adaptés à l'analyse des médias sociaux.

## 9.3 Application à Galaxy2

Le second jeu de données investigué provient de Galaxy2, un petit réseau social basé sur TOR. Dans cette section, nous présentons TOR et Galaxy2, puis parcourons les trois niveaux d'analyse, sur les textes, les comptes et les relations sociales. Nous montrons ainsi que SARTN est capable de traiter d'autres types de données, et notamment d'exploiter la variété des actions et publications présentes dans Galaxy2.

### 9.3.1 Présentation de Galaxy2

La spécificité de Galaxy2 repose dans son inaccessibilité depuis un navigateur web classique : il faut impérativement se connecter au DarkNet, et plus précisément au réseau TOR.

#### 9.3.1.1 Le réseau TOR

Partiellement développé par la DARPA (agence états-unienne de l'armement) dans les années 1990, TOR, *The Onion Router*, a été mis à disposition du public en 2002. Visant à apporter l'anonymat au flot de données, il est basé sur un protocole de chiffrement et de routage en oignon, nommé « onion » : les contenus sont chiffrés puis transitent par un autre nœud du réseau TOR, où ils sont de nouveau chiffrés puis envoyé à un autre relais. Les paquets sont susceptibles de ne pas tous transiter par les mêmes nœuds, cachant ainsi les contenus vis-à-vis de l'infrastructure du réseau Internet.

TOR est utilisé dans deux cas de figure : soit pour accéder anonymement au Web classique, soit pour accéder au *DarkNet* ou *Dark Web*. En effet, des sites Web sont accessibles uniquement via le protocole *onion*, afin d'en protéger l'hôte. Des exemples célèbres incluent Wikileaks, permettant de diminuer le niveau de risque couru par les lanceurs d'alerte, ou encore *the silk road*, une plate-forme de e-commerce illégale désormais démantelée.

Ainsi, chacun peut utiliser le réseau TOR et profiter d'un certain anonymat [3]. TOR aide à passer outre la censure, et est utilisé par des journalistes, activistes politiques, ou autres : nul besoin de se justifier. Malheureusement, TOR est aussi exploité pour héberger des contenus illégaux, dont des forums de hackers, des sites de vente de stupéfiants, des contenus pornographiques [4].

Cependant, à cause de sa complexité perçue d'utilisation, et surtout des bénéfices ressentis par chacun de connecter son profil Internet à sa vie réelle (shopping, recommandations culturelles, gestion d'actifs bancaires, relations sociales pour ne citer que ceux-ci), la quantité d'internautes sur le *DarNet* n'est pas très grande. De plus, certains états déploient des lois, règles et pares-feu qui limitent fortement son expansion. La figure 9.22 (en anglais, issue des travaux de *Stefano De Sabbata* [5]) illustre l'utilisation mondiale de TOR en 2014, attribuant une taille aux pays en fonction du nombre total d'utilisateurs. L'Europe occidentale concentre la majorité d'entre eux ; la Russie et le Moyen-Orient sont aussi très présents. Sur cette carte, il manque un géant en Asie : la république populaire de Chine n'approuve pas ce genre d'outils.

#### 9.3.1.2 Galaxy2 : histoire et caractéristiques principales

Fondé en 2015 après l'interruption d'un précédent réseau social hébergé sur TOR (« Galaxy », premier du nom), *Galaxy2* est basé sur un framework libre nommé *elgg*, qui facilite la construction de sites sociaux : il comprend des fonctionnalités telles que des outils de création de profils, de pages personnelles, de publication de contenus, en couplant une base de données et un serveur Web.

3. L'utilisation seule de TOR ne garantit pas l'anonymat.

4. Cela n'empêche pas la loi de s'y appliquer, plus difficilement peut-être. `https://www.fbi.gov/news/stories/playpen-creator-sentenced-to-30-years`

5. Carte publiée sur son blog, `https://stefanodesabbata.com/2014/06/09/tor/`

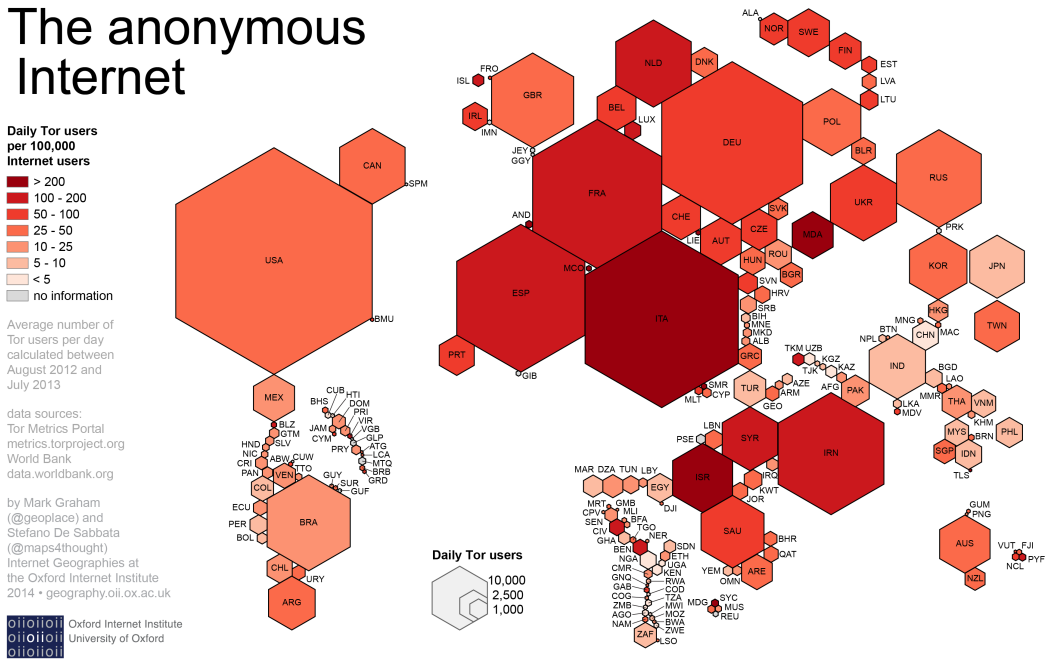



FIGURE 9.22 – Géographies de TOR, *Stefano De Sabbata*

Selon son fondateur, *Lameth*, le serveur est tombé, sans qu'aucune sauvegarde n'ait été réalisée. Ainsi le service est indisponible depuis fin octobre 2017 [6]. À l'adresse habituelle, seul un message d'excuses accueille le visiteur ; ce message est reproduit en annexe B.3 à titre d'anecdote.

Les principales fonctionnalités de Galaxy2 incluent *TheWire*, un espace de micro-blogging ; des pages plus consistantes dédiées aux blogs, sondages et groupes ; et enfin, un partage d'images et de fichiers. En lien avec l'anonymité introduite par le réseau TOR, les utilisateurs ont tendance à ne pas confier leurs données personnelles : les pseudonymes ne ressemblent pas à de vrais noms ; les photos n'illustrent pas les vacances de chacun. En conséquence, à l'exception des messages directs privés, le reste des données n'est absolument pas protégé : toutes les publications, connexions d'amitiés, images et actions accomplies par les utilisateurs inscrits étaient directement accessibles sur la page visible en figure 9.23, durant les presque trois ans de fonctionnement de Galaxy2.

Ainsi, la collecte que nous avons réalisée couvre un intervalle de temps allant de la création de Galaxy2, le 9 janvier 2015, jusqu'à la date de fin de collecte, le 22 septembre 2017. Il était prévu d'actualiser le corpus régulièrement, mais la chute du service est un obstacle incontournable. Le corpus contient toutes les traces d'activité publiquement disponibles : la publication d'un message ou l'abonnement à un compte en font partie ; le fait de se connecter ou de lire des messages, non. L'information de publication d'une image est retenue, mais pas l'image elle-même.

#### 9.3.1.3 Description du corpus : types d'actions réalisées

La table 9.7 présente le nombre d'actions accomplies, réparties par type d'action, puis par ordre croissant, sur la période couverte par la collecte. Les actions sont ici regroupées pour faire ressortir leur diversité : en premier, les commentaires. En effet, les utilisateurs peuvent ajouter quelques mots pour réagir aux actions des autres, qu'il s'agisse de publications de photos, de fichiers, d'une nouvelle page ou d'un sondage ; et bien sûr aux billets de blog. Les trois activités de type « *commented on* » correspondent à un artefact lors du commentaire d'un billet de blog ; les autres catégories sont heureusement plus explicites.

Un second groupe d'action rassemble la publication d'images et de fichiers. Ceux-ci sont voués au partage, et consistent fréquemment en des sortes d'affiches de publicité / propagande de la mou-

6. https ://socialmediaalternatives.org/archive/collections/show/10


FIGURE 9.23 – Apparence de Galaxy2

vance *Anonymous*. Les philosophies en vogue sur TOR et Galaxy2 n'encouragent vraiment pas à la distribution de photos personnelles, ou de vacances ; aussi la quantité totale de photos publiées est très faible (pour un total d'environ 600 images sur l'ensemble du réseau). Les *avatars* sont en fait des changements de photos de profil (mais qui ne sont pas comptabilisées comme des publications de photos) ; ici encore ce n'est pas la fonctionnalité la plus utilisée sur Galaxy2.

La création de pages, groupes et sondages, ainsi que leurs modifications (ajout de nouveaux éléments, des « *topics* » au sein des groupes) est un des piliers de la vie du réseau. Ces fonctionnalités facilitent une discussion continue sur un sujet donné, parmi un ensemble d'abonnés ; elles servent de support pour les actions de votes, réponses et publication sur ces pages / groupes.

Les fonctions de microblogging étaient prédominantes sur Galaxy2, avec 29 000 billets postés sur *TheWire*, le fil public du réseau. La taille du message n'était pas limitée ; si quelques messages sont effectivement longs, l'immense majorité ne contient qu'une ou deux phrases.

Finalement, le dernier type d'actions est constitué des informations relationnelles : notifications de création de compte (« X *joined the site* »), établissement de liens d'abonnements entre un compte et une page, un groupe ou un autre compte (lien « d'amitié »).

### 9.3.2 Analyse du texte : répartitions entre thématiques et sentiments

Cette section présente les thématiques détectées dans le corpus des publications de Galaxy2, et procède dans un second temps à une analyse des messages et des utilisateurs à partir du sentiment de leurs émissions.

#### 9.3.2.1 LDA pour la détection de thématiques

La méthode LDA (*Latent Dirichlet Allocation*) [Blei et al., 2003] est communément paramétrée avec un grand nombre de clusters de documents : entre 200 et 500 selon les recommandations pour des corpus de textes longs [7]. Ne disposant pas d'annotations, il est difficile de trouver un paramètre optimal. L'un des facteurs limitant, la complexité du calcul, dépend d'une part du nombre de documents, et d'autre part du nombre de thématiques choisi. Au vu de la quantité de documents dont dispose le corpus, et comme un compromis entre temps de calcul et représentativité des thématiques, un nombre de 40 thèmes est retenu. Comparé à d'autres valeurs plus importantes, ce paramétrage limite le nombre de thématiques trop proches l'une de l'autre.

7. http ://radimrehurek.com/gensim/models/ldamodel.html

TABLEAU 9.7 – Description des types d'actions présentes dans le corpus

| | Type d'action | Quantité |
|---|---|---|
| Commentaires | commented on | 3 |
| | commented on the album | 149 |
| | commented on the file | 170 |
| | commented on a page titled | 192 |
| | commented on a bookmark | 239 |
| | commented on the photo | 375 |
| | commented on the poll | 672 |
| | commented on the blog | 4 032 |
| Fichiers | uploaded the file | 323 |
| | created a new photo album | 381 |
| | added < $some$ > photo(s) | 516 |
| | has a new avatar | 1 442 |
| Création | created a page | 109 |
| | created a poll | 113 |
| | added a new discussion topic | 451 |
| | created the group | 548 |
| | published a blog post | 1 944 |
| Discussions | voted on the poll | 1 328 |
| | replied on the discussion topic | 1 340 |
| Microblogging | posted on < $a\ wall$ > | 606 |
| | posted to < T$he$ W$ire$ > | 29 210 |
| Connections | bookmarked | 541 |
| | joined the site | 19 233 |
| | joined the group | 26 566 |
| | is now a friend with | 61 027 |

La figure 9.24 est fournie par un outil, *pyLDAvis*, proposé par [Sievert and Shirley, 2014], qui permet de naviguer parmi les 40 thématiques détectées au travers des fréquences (conditionnelles ou non) de mots dans un cluster donné (en rouge), comparées aux fréquences des mots dans l'ensemble du corpus (en bleu), dans le panneau à droite. Sur les vecteurs $tf.idf$ des clusters de documents, une analyse en composantes principales (ACP) est appliquée, afin d'attribuer une position aux thématiques sur un espace en 2D, sur la partie gauche du diagramme. L'aire d'un cercle est proportionnelle à la quantité de documents regroupés dans la thématique qu'il représente.

Grâce à cet outil, la distribution des thématiques est visualisable, illustrant ainsi 30 000 documents en une image, tout en permettant de les explorer par la fréquence des mots dans chacun des groupes de documents.

#### 9.3.2.2 Exploitation du sentiment

Outre la thématique, chaque texte est analysé pour en dégager le sentiment, représentant la polarité globale du texte, perçue par le lecteur. Nous utilisons ici l'outil Vader [Hutto and Gilbert, 2014], que nous avons présenté en chapitre 2 ; l'objectif est ici d'obtenir une vue globale des sentiments émis. Ainsi, la figure 9.25 compare les polarités des textes contenant les mots-clés suivants : *onion,* TOR*, Snowden, Trump* et *Syria* [8].

Pour représenter le sentiment mesuré sur un ensemble de textes, la moyenne n'est pas l'unique modalité. Pour compléter le « Sentiment Net », illustré dans l'étude précédente en figure 9.1, nous utilisons les boîtes à moustache en figure 9.25, affichant les médianes et les quartiles. *Onion* est

8. L'analyse est réalisée indépendamment de la casse : majuscules et minuscules ne sont pas différenciées ici.

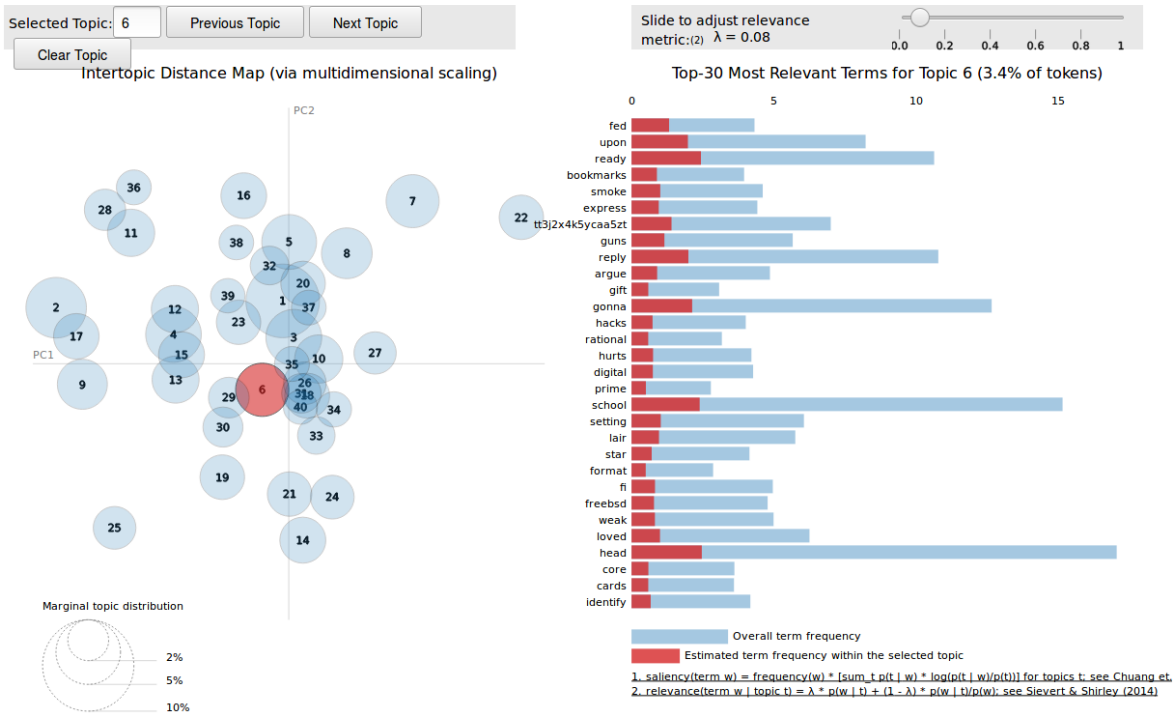



FIGURE 9.24 – Visualisation des thématiques

globalement bien perçu, avec des messages positifs. Edward Snowden semble bénéficier d'une bonne réputation, tandis que les membres de Galaxy2 sont plus mitigés envers Donald Trump. Finalement, la polarité est plus négative lorsque la Syrie est évoquée : ces messages sont susceptibles d'évoquer le conflit actuel.

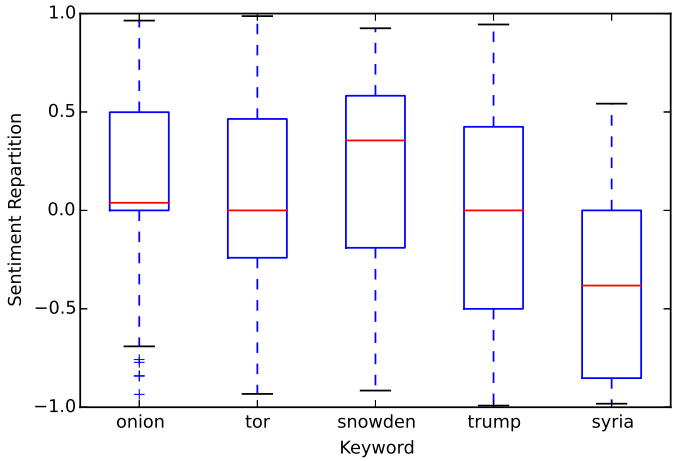



FIGURE 9.25 – Répartition du sentiment pour quelques mots-clés

Les données issues du sentiment servent aussi à caractériser l'activité d'un utilisateur. À titre d'exemple, la figure 9.26 expose les polarités des messages de deux utilisateurs (très) actifs : *XL33t* et *Fenris*. Sur chacun des deux graphiques, un artefact est visible en abscisse 0, représentant les messages pour lesquels aucune polarité n'a pu être extraite (aucun mot « chargé » reconnu). Malgré cela, la répartition des polarités suit une tendance : dans le premier graphique, *XL33t* semble émettre surtout des messages très positifs. Le second auteur, *Fenris*, couvre l'ensemble du domaine des possibles, y compris le négatif. Grâce à cette visualisation, il est possible de se faire une idée assez précise de

cet aspect du comportement des utilisateurs.

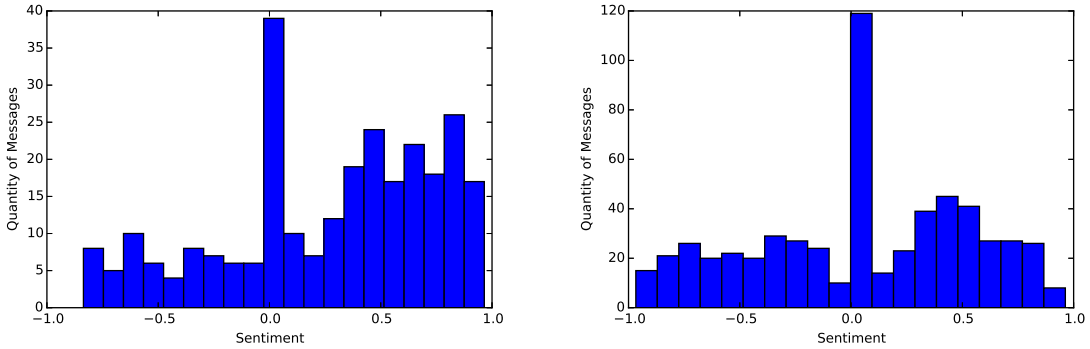



FIGURE 9.26 – Visualisation de sentiment émis par deux utilisateurs : XL33t et Fenris

Ces visualisations complètent le tableau de bord, décrit lors de l'étude du corpus *KevRandTweets* et illustré en figure 9.2 ; nous l'omettons pour éviter la redondance. L'ensemble de ces traitements et modules de visualisation interactive permettent de suivre les thématiques, entités et sentiments respectifs, constituant les opinions diffusées sur le réseau ; cela correspond au critère 1 énoncé au début de ce chapitre.

### 9.3.3 Représentation du réseau social par des graphes

Un réseau social numérique propose de nombreuses fonctionnalités pour créer des liens entre les comptes ; l'usage qui en est fait par les utilisateurs rajoute des interprétations possibles à ces liens. Ainsi, nous proposons de représenter le réseau social par trois graphes, $G_I$, $G_F$ et $G_\Omega$, chacun représentant un type de lien social.

#### 9.3.3.1 L'interaction par les mentions

Une pratique commune sur les réseaux sociaux consiste à mentionner des comptes, en plaçant une arobase « @ » devant le pseudonyme de l'autre, dans le texte des messages. Au début, il ne s'agissait que d'un comportement social sans fonctionnalité particulière de la plate-forme ; dorénavant la plupart des réseaux sociaux numériques reconnaissent ces mentions, y associent un lien direct vers le profil du compte mentionné, et gèrent l'envoi éventuel de notifications.

Puisque Galaxy2 est basé sur le framework *elgg*, qui est relativement ancien, une telle fonctionnalité n'est pas implémentée ; cependant ce comportement est bel et bien en vogue chez les utilisateurs de Galaxy2. Afin de les exploiter, un module d'extraction de mentions parcourt chacune des publications à la recherche de signes « @ » suivis de pseudonymes, générant ainsi une liste d'interactions directes entre comptes.

À partir de l'ensemble ME des mentions entre utilisateurs, un graphe d'interaction par la mention, $G_I = (V, E_{ME})$, est construit. V est l'ensemble des comptes, qu'ils soient auteurs de mentions, ou comptes mentionnés. Les arcs $e \in E_{ME}$ relient l'auteur du message aux utilisateurs mentionnés dans le message. Ce graphe $G_I$ est composé de 968 nœuds, reliés par 2 342 arcs, pour un total de 5 481 mentions : les arcs sont pondérés par le nombre de mentions de même auteur et de même destinataire. C'est peu par rapport aux 20 000 comptes présents sur Galaxy2, mais cohérent : l'étude des comportements des comptes a révélé que la majorité des comptes sont passifs (et n'utilisent donc pas les mentions).

#### 9.3.3.2 Les amitiés

La fonctionnalité qui a fait la renommée du Web 2.0 repose sur les connexions entre « amis ». Sur Galaxy2, un compte revendique être l'ami d'un autre ; une fois validée par l'autre, la relation est établie, réciproque entre les comptes. L'action d'ajout d'un compte parmi sa liste d'amis fait partie des actions collectées, et permet de construire le graphe des amitiés ou abonnements, $G_F$, dont les nœuds sont les comptes utilisateurs, reliés par les arcs d'amitié. Dans le corpus collecté, $G_F$ contient 7 356 nœuds, reliés par 60 860 arcs. Le nombre de nœuds est à mettre en relation avec la quantité de comptes créés sur Galaxy2 (19 233 au total) : il est fréquent de créer un compte, sans l'utiliser ni même effectuer une demande d'ami.

#### 9.3.3.3 Partage d'objets sociaux

Les comptes utilisateurs sont aussi reliés par les objets (images, pages, etc) qu'ils partagent. Sur Galaxy2, certaines fonctionnalités viennent enrichir la liste des objets sociaux (notion d'album photo, sondages ou votes, pages de discussion) ; d'autre part, l'émission de supports multimédia ou de liens Web n'est pas aussi facilité que sur l'étude de cas précédente. En conséquence, nous retenons trois types d'objets : les **photos** (et les commentaires suscités), les **pages de vote** (*created a poll* : puis l'acte de vote, et les commentaires liés), et enfin les **pages de discussion** (*discussion topic* : création et commentaires) pour constituer trois graphes d'échange, et finalement les assembler, résultant en $G_\Omega$, le graphe de partage d'objets sociaux.

La notion de partage intervient en deux temps sur Galaxy2 :

1. un compte, *l'émetteur*, crée l'objet, au moyen des fonctionnalités fournies : création d'un album photo, ajout d'une ou plusieurs images à un album, création d'une page de vote, ou d'une page de discussion sur un sujet donné ;
2. des comptes utilisateurs, *les partageurs*, réagissent au contenu ajouté par l'émetteur : ils commentent les photos, prennent part aux votes et répondent sur les pages de discussion.

La construction des graphes $G_{photo}$ (respectivement, $G_{vote}$, $G_{disc}$) est réalisée comme suit : chaque photo (respectivement, sondage ou discussion) est directement reliée à son unique *émetteur* ; ainsi, à chaque réaction, un arc est ajouté du *partageur* vers l'*émetteur* de la photo en question. Le graphe final, $G_\omega$, est l'union de $G_{photo}$, $G_{vote}$ et $G_{disc}$. Il contient 1 092 nœuds et 2 064 arcs.

Ces graphes sont nécessaires pour poursuivre l'analyse de Galaxy2, dans un premier temps par la détection d'acteurs-clés et la caractérisation du comportement des comptes utilisateurs ; puis dans un second temps, pour identifier les communautés thématiques présentes sur certains de ces graphes. Ils diffèrent des graphes construits à partir de *KevRandTweets*. L'usage de la mention est ici moins répandu, les liens d'amitié sont présents, et les types d'objets sociaux partagés ne sont pas les mêmes, dépendant des usages en vogue sur Galaxy2.

### 9.3.4 Comptes influents et types de comportement

Après cette phase d'étude des messages publiés, nous nous penchons sur les acteurs-clés de Galaxy2, puis sur les comportements-types.

#### 9.3.4.1 Top-5 des acteurs-clés

L'influence est un concept mesurable au travers d'une variété de prismes. Au premier regard sur un réseau social, il est fréquent d'y chercher des « acteurs-clés » ; nous proposons un « top5 ». La table 9.8 compare les comptes les plus connectés, c'est-à-dire qui ont le plus grand nombre d'amis, notés dans la colonne #Friends, ainsi que leur score de *popularité*, pour lequel un score de 1 correspond à un total de 10 000 amis ; les plus mentionnés, notés dans la colonne #Mentions ; les utilisateurs de référence selon notre score d'influence introduit en chapitre 6 (le score PageRank calculé sur le graphe

des mentions), dans la colonne *Influence* ; et les comptes les plus actifs, réalisant le plus d'actions sur Galaxy2.

En quelques mots, nous rappelons l'idée derrière le calcul d'un score d'influence. Inspiré de la littérature, où un score d'influence est souvent extrait du graphe des abonnements [Kwak et al., 2010, Lee et al., 2010], nous préférons calculer le nôtre sur une relation plus active : les mentions (présence du pseudonyme précédé d'une arobase dans le texte du message). L'intuition est la suivante : une mention donne de la valeur, du capital social, à l'utilisateur mentionné. Cette valeur est d'autant plus élevée si l'émetteur de la mention dispose lui-même d'un certain capital social. L'algorithme PageRank suit aussi cette intuition, et correspond à cette représentation de l'influence. Les valeurs de scores PageRank ne portent pas de sens intrinsèque : ici, l'*influence* est le **classement** résultant des scores PageRank de chaque nœud-compte, dans le graphe des mentions $G_M$, qui reflète une partie des interactions entre comptes. L'absence de fonctionnalité de « plussoiement » (ou de retweet) empêche le calcul de l'expertise.

TABLEAU 9.8 – Top5 des comptes utilisateurs selon les amitiés, mentions et actions

| #Friends | | Popularité | #Mentions | | Rang d'Influence | #Actions | |
|---|---|---|---|---|---|---|---|
| XsyntaX | 8 187 | 0.98 | XsyntaX | 514 | Oxyy | XsyntaX | 7 974 |
| prozac | 3 182 | 0.88 | Oxyy | 164 | XsyntaX | kheper | 1 846 |
| Spooky | 3 012 | 0.87 | Lameth | 121 | Lameth | prozac | 1 783 |
| xl33t | 2 140 | 0.83 | Fenris | 119 | ChatTor | ChatTor | 1 583 |
| kheper | 1 922 | 0.82 | cpnnemo | 96 | MahaKali | Spooky | 1 569 |

Quelques mots pour accompagner la table 9.8 : *XsyntaX* dépasse largement tous les autres inscrits, avec 7 974 actions accomplies durant la période analysée ; le second, *kheper*, n'en a réalisé « que » 1 846. Ces quantités, impressionnantes, sont toutefois très nettement inférieures à ce qui est constaté sur Twitter, où de nombreux robots se distinguent par leur hyper-activité. *Lameth* est le fondateur de ce réseau social, et c'est en tant que tel qu'il est souvent mentionné, que ce soit en remerciement ou pour des suggestions d'amélioration du service. *ChatTor* fournit un service externe de chat en ligne via TOR ; ce compte fait la promotion du service, et en publie les actualités.

Bien que certains noms soient présents dans le Top5 sur chaque aspect de l'influence, ils ne sont pas tous également influents. *Lameth* est énormément mentionné, de par son statut d'hôte. Cependant, il ne souhaitait pas être l'ami de chaque utilisateur de Galaxy2, ni être le plus actif. *Xl33t* est ami avec de nombreux comptes, mais n'est pas autant mentionné que les autres acteurs-clés. L'influence prend ainsi de nombreuses formes, mais seuls les symptômes sont mesurables. Selon l'application recherchée, l'utilisateur du système se penchera sur l'indicateur d'influence le plus adapté, ou les mettra en relation pour comparer les acteurs-clés découverts, ce qui correspond au critère 2.

#### 9.3.4.2 Types de comportement

Afin de caractériser les comportements adoptés par la population de comptes sur Galaxy2, nous répartissons les données le long de quelques aspects.

La *Biographie* contient les informations liées à l'identité de l'utilisateur : pseudonyme, ID, date de création de compte. Le *Style* explique comment l'utilisateur écrit, par la longueur des messages, la quantité de ponctuation, les thématiques d'intérêt et la polarité de sentiment.

L'aspect *Social* détaille le nombre d'amis, de mentions reçues et émises, ainsi que le score d'influence du compte. La dimension *Media* regroupe les caractéristiques expliquant le type d'actions accomplies, et la quantité d'objets sociaux postés (que ce soient des photos, URLs, etc). Finalement, l'aspect *Temporel* couvre le rythme de publication du compte : nombre moyen de « tweets » (ici, des billets sur *TheWire*) par jour, et nombre moyen d'actions par jour.

L'ensemble de ces caractéristiques numériques permet de répartir les utilisateurs entre quelques clusters, dont le nombre réduit (habituellement entre 4 et 6 ; ici, quatre clusters suffisent car la population totale est faible) facilite la compréhension des classes obtenues.

Le processus suivi est expliqué en chapitre 6 ; nous le rappelons brièvement ici. Tout d'abord un nettoyage des données est nécessaire afin d'éviter les divisions par zéro (pour les quantités moyennes par exemple) ou encore les distributions exponentielles de certaines caractéristiques (auquel cas, la caractéristique est remplacée par sa valeur logarithmique). Le jeu de données est normalisé pour être centré (moyenne à zéro) et réduit (écart-type à 1). Ensuite, une ACP réduit la dimension du problème, résumant 34 dimensions en seulement 5, tout en conservant la majeure partie de la variance. L'importance de chacune des caractéristique auprès des 5 dimensions de l'ACP est présentée en annexe B.4. Finalement, l'algorithme des k-moyennes découvre les groupes de comportements similaires ; nous recommandons de fixer le paramètre $k$ au regard de l'optimisation de la mesure de Calinski-Harabasz [Caliński and Harabasz, 1974], comme illustré dans l'étude de cas précédente, en figure 9.4. Selon l'application, l'utilisateur de SARTN peut toutefois retenir une valeur différente, résultant en une répartition correspondant mieux au cas d'usage. Nous choisissons ici empiriquement de rechercher quatre profils-types.

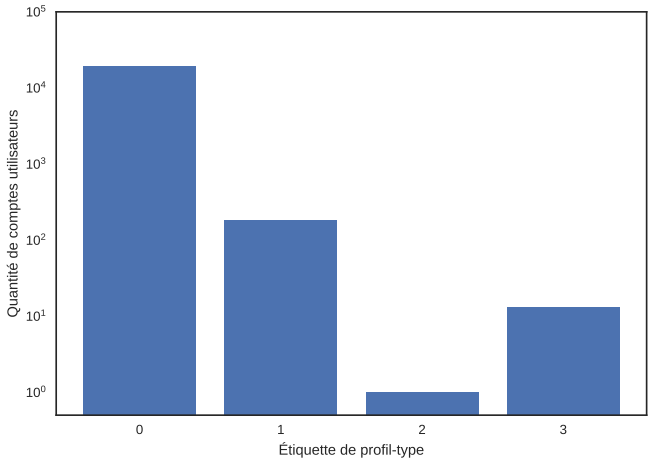



FIGURE 9.27 – Répartition des profils des utilisateurs

La figure 9.27 montre la répartition des utilisateurs entre les quatre types de profils comportementaux. Le premier type, noté 0, est de loin le plus commun : il inclut plus de 90% des comptes. En effet, la majorité des comptes n'accomplissent qu'une ou deux actions après leur création, et sont parfois oubliés, inactifs. Un second type, le profil 1, caractérise les utilisateurs réguliers actifs, même s'il ne s'agit pas des comptes « centraux ». Finalement, le profil 2 ainsi que l'unique compte de profil 3 sont les comptes importants, très actifs, qui produisent la plupart des contenus émis sur le réseau. Une répartition similaire (mesurée uniquement en nombre de messages publiés et nombre d'amis) a déjà été observée sur d'autres plate-formes comme Twitter [Kwak et al., 2010].

Afin de visualiser la forme des clusters de profils, la figure 9.28 représente chacun des 19 177 comptes utilisateurs de Galaxy2, projetés sur les deux premières dimensions de l'ACP. Ces dimensions sont des combinaisons linéaires des caractéristiques comportementales, et donc sont assez abstraites. Cependant, les positions proches de (0,0) sont liées à des niveaux très faibles d'activité (le cluster (0) étant le plus peuplé). En bleu, le profil 1 montre les différentes manières d'être un utilisateur normal, « actif mais pas trop ». Le profil 2, en vert, est constitué d'un unique compte, *Spooky*, qui réalise en moyenne 2,5 actions par jour. Finalement, le profil 3, en violet, est très dispersé, mais correspond à un haut niveau de publication, que ce soit par nombre de billets sur *TheWire*, par des commentaires, ou par la création de pages.

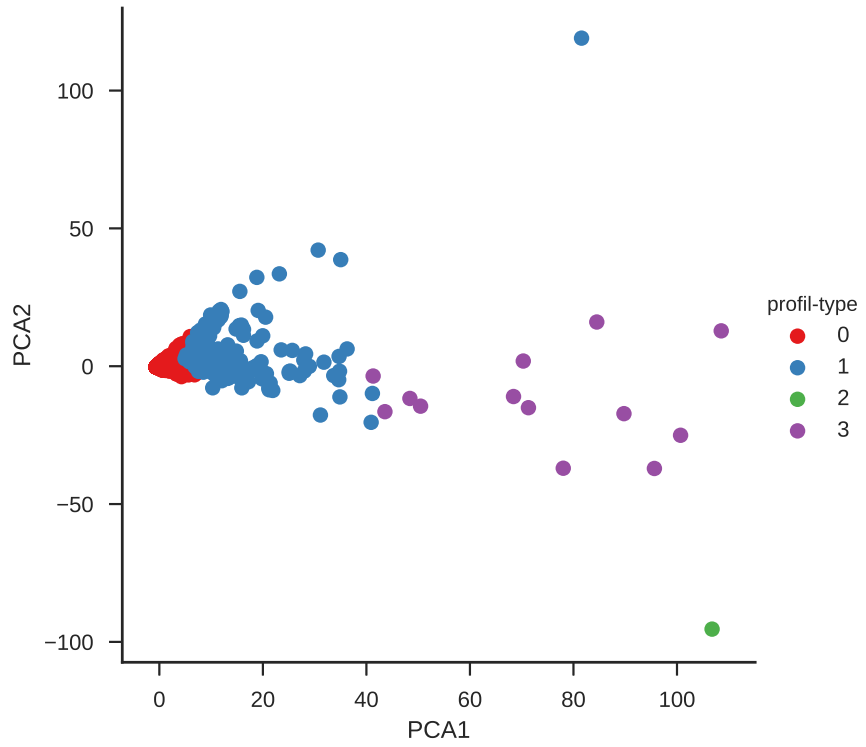



FIGURE 9.28 – Visualisation des clusters d'utilisateurs dans la projection de l'ACP

#### 9.3.4.3 Rôle-type des comptes concernant les objets sociaux

L'étude du graphe $G_\Omega$ est propice pour quantifier les échanges d'objets tels que les images, les sondages et les pages de discussion. L'application de l'algorithme de détection de rôles-types sur ce graphe résulte en une distribution particulièrement déséquilibrée : le rôle le plus actif, rôle 0, ne regroupe que 6 comptes, contre 795 pour le rôle 1, 174 pour le rôle 2 et 117 pour le rôle 3.

La figure 9.29 quantifie ces rôles-types, en montrant la distribution des degrés. Celle-ci est ici tronquée : le degré maximum atteint par rôle 0 vaut 350, unique point très éloigné au-dessus du reste de la population. Le rôle 1 constitue la majorité des *partageurs*, qui sont globalement peu actifs ; il en va de même pour le rôle 2. Le rôle 3, en rouge, affiche de degrés significativement supérieurs, qui « consomment » souvent de tels contenus.

Cette étape d'analyse permet de considérer, par le **profil-type**, le comportement adopté par l'auteur d'un message, et ainsi différencier s'il provient d'un compte bien établi, influent, ou bien si le message provient d'un compte habituellement « muet », c'est-à-dire très peu actif. Par le **rôle-type**, cela fournit des informations sur la position sociale adoptée par le compte. Cette capacité remplit l'exigence énoncée en critère 3, concernant la caractérisation des comptes.

### 9.3.5 Détection et caractérisation des communautés

Le jeu de données provenant de Galaxy2 contient plusieurs informations susceptibles de servir de support à un graphe, parmi lesquelles nous retenons les **mentions** (un type spécifique d'interaction), comme lors de l'étude de cas portant sur Twitter, ainsi que les abonnements ou **liens d'amitié**, et les **partages d'objets**, dont les images, les sondages et les pages de discussion.

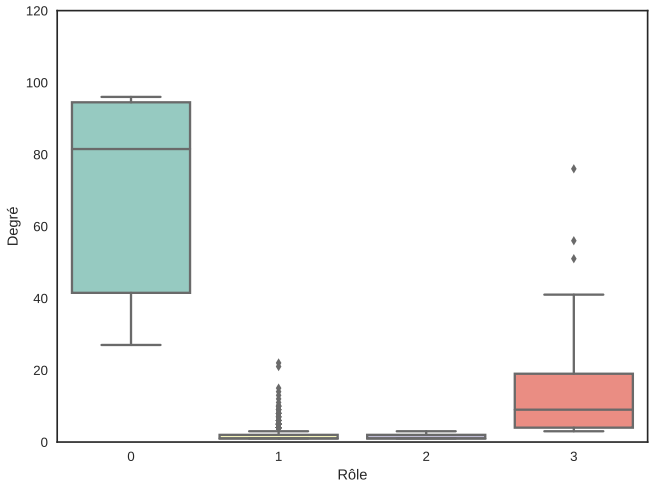



FIGURE 9.29 – Degré selon le rôle du nœud dans $G_\Omega$

#### 9.3.5.1 Interaction par la mention

Le support d'interaction directe retenu dans cette étude de cas repose sur les *mentions*, et permet de construire le graphe $G_I$. L'algorithme de détection de communautés Louvain [Blondel et al., 2008] y détecte 31 communautés, parmi lesquelles seules 11 contiennent plus que 3 membres ; cette détection est la première étape nécessaire pour satisfaire le critère 5. En figure 9.30, les nœuds représentent ces groupes ; des arcs signalent des liens inter-communautés.

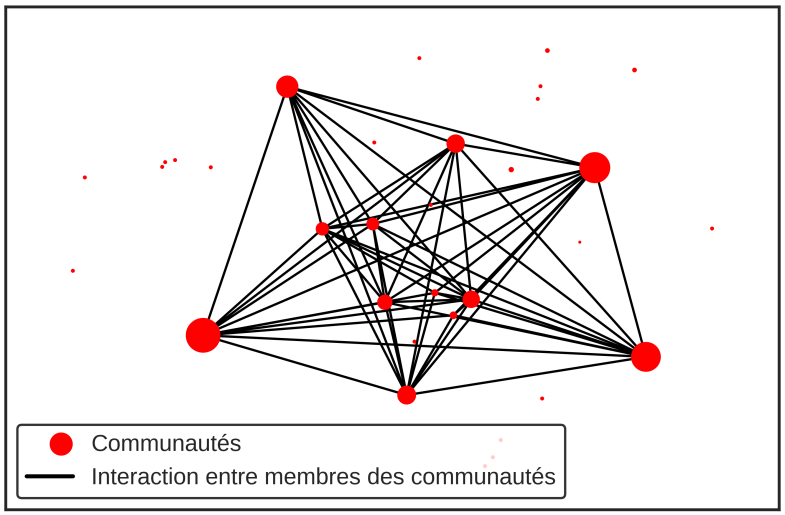



FIGURE 9.30 – Visualisation des liens entre communautés issues du graphe des interactions $G_I$

Cette visualisation est intéressante, et permet de représenter l'ensemble d'un réseau. Cependant, les liens ne sont pas différenciés, selon qu'il y ait eu un ensemble de mentions, ou une seule ; enfin, les communautés ne sont qualifiées que par leur taille : il y manque des informations sur leur cohésion interne, structurelle et thématique.

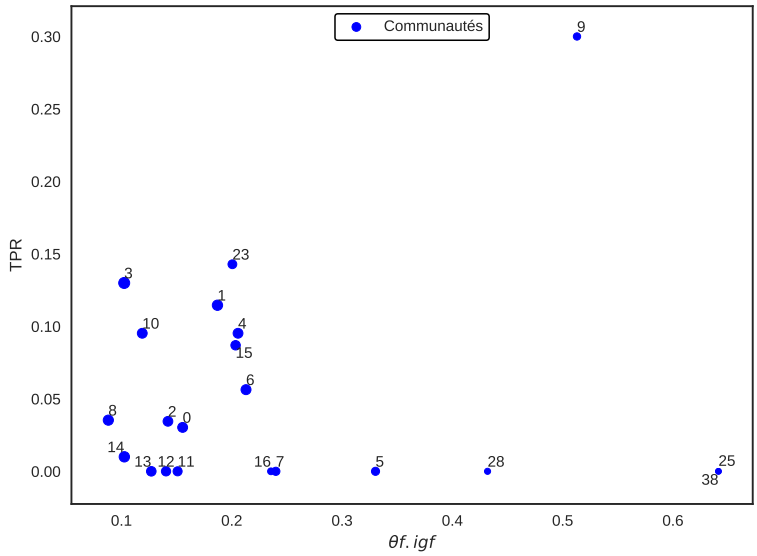



FIGURE 9.31 – Comparaison des $\theta f.igf$ et de TPR des communautés de $G_I$

La figure 9.31 vient pallier ce manque, en représentant en abscisses $\theta f.igf$, la pertinence thématique des communautés détectés, et en ordonnées le ratio de triangles TPR, c'est-à-dire leur cohésion structurelle. Ces communautés sont très faiblement cohésives : le meilleur TPR = 0.184 ne correspond pas à un groupe dont les membres se connaissent les uns les autres. D'autre part, les scores thématiques sont particulièrement bas, ne permettant pas de caractériser un groupe par la thématique principale qu'il aborde.

Le faible nombre de mentions détectées, conséquence de l'absence de fonctionnalité dédiée, diminue dans cette étude de cas l'importance du graphe des interactions. D'autres supports de relations entre comptes sont accessibles pour représenter l'activité du réseau social d'une autre façon.

#### 9.3.5.2 Le graphe des amitiés

La relation d'amitié est répandue sur Galaxy2, et fait partie du jeu de données collecté. Elle sert de support à la construction du graphe d'amitié ou d'abonnement, noté $G_F$. Sur ce graphe, les communautés sont révélées par l'algorithme Louvain. Un total de 32 communautés sont obtenues, dont 11 incluent strictement plus que 3 utilisateurs. La figure 9.32 montre la distribution des communautés sur deux mesures : l'axe des abscisses représente la conductance, c'est-à-dire la proportion d'arcs liant une communauté à une autre (en relation au nombre d'arcs internes à la communauté). La conductance évalue dans quelle mesure une communauté est liée à son environnement, montrant soit son influence, soit son isolement. L'axe des ordonnées représente la représentativité ρ des groupes, qui mesure la présence du groupe sur sa thématique majoritaire. Un ρ élevé signifie que tous les utilisateurs actifs sur une thématique donnée sont membres du groupe.

Les groupes présentent tous des conductances élevées en figure 9.32, ce qui s'explique par la nature du graphe des abonnements : ceux-ci sont nombreux et réciproques, reliant presque tous les comptes en un ensemble connexe. Les groupes en bas à droite sont actifs sur des thématiques largement distribuées dans le réseau, limitant leur impact relatif.

En figure 9.32, le groupe n°2 attire l'attention. De conductance 0,74 et de ρ = 0.054, cette communauté de 125 membres est nettement moins reliée au reste du réseau que ses paires. Pour l'analyser plus en détail, ses scores thématiques, introduits en chapitre 7 et liés au critère 6, donnent des clés d'interprétation. $\xi_u$ = 0.064 signifie qu'un maximum de 6,4% de ses membres ont été actifs sur une

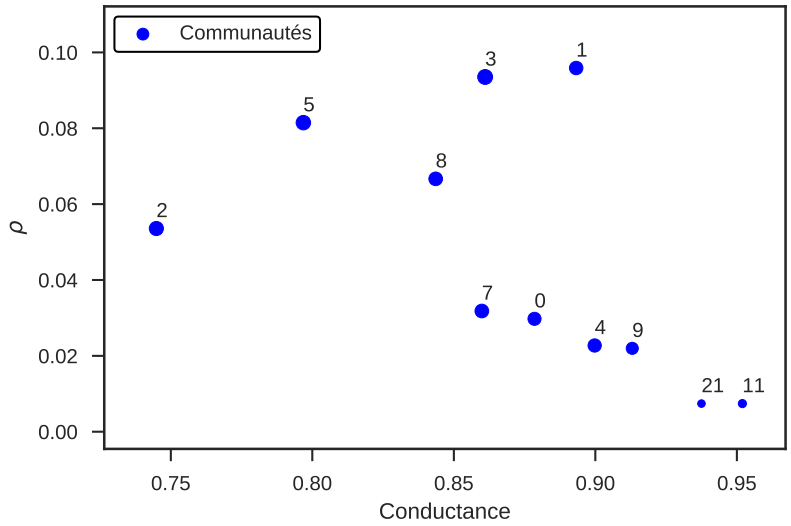



FIGURE 9.32 – Comparaison de la conductance et de la représentativité de communautés issues de $G_F$

même thématique ; cependant, comparée à la durée de la collecte, cet état de fait n'est pas surprenant. En vérifiant les dates des textes de cette thématique principale, nous pouvons la décomposer, selon l'époque, en deux sous-thèmes. Le premier réfère un un groupe de discussion à caractère pornographique ; le second à l'utilisation d'un protocole de mail en pair-à-pair, à travers TOR.

Ainsi, la relation d'amitié est plus propice que les mentions pour la détection de communautés intéressantes, dont l'exploration est facilitée par les mesures topologiques et thématiques. La combinaison de ces deux relations permet de mettre en relief un groupe détecté sur un graphe, grâce aux informations provenant du second.

#### 9.3.5.3 Comparaison des amitiés et des interactions par les mentions

La figure 9.33 illustre une petite communauté parmi celles détectées sur le graphe des interactions ; elle porte le n°11 en figure 9.31. Nous avons choisi ce groupe car il affiche un score de cohésion thématique respectable $\xi = 0.40$, ainsi qu'une petite taille (facilitant l'affichage). Les nœuds en rouge font partie du groupe détecté ; les nœuds bleus sont au-delà de la frontière du groupe dans le graphe $G_I$. Le compte central, nommé *Bishop*, est ciblé par quelques mentions, notamment de la part d'un compte audacieux *Nishikino_Maki*[9]. D'un regard, nous voyons que ce groupe est centré sur *Bishop*, qui n'est pourtant pas le plus actif dans la vie du groupe (il ne mentionne pas une seule fois les autres membres du groupe).

Cette vue d'une communauté illustre les liens entre un compte et le reste du réseau, accompagnant l'exploration de Galaxy2 en parcourant les liens d'interaction entre les comptes. Cependant, comme le graphe des interactions ici exploité est construit à partir d'un usage social, c'est-à-dire les mentions par l'utilisation de l'arobase, il est probable que l'outil de détection des mentions a manqué reconnaître des interactions écrites différemment ; le graphe peut être incomplet.

Le groupe concerné, contenant 5 membres, présente une faible densité interne ($d_{int} = 0,4$) au regard de sa dimension. Cette faiblesse est rappelée par l'absence de triangles (TPR = 0). Parmi les 5 membres, deux sont actifs sur le même thème, permettant d'atteindre un $\xi = 0.4$ ; ce thème se résume à des recommandations d'URLs en *.onion*, sans relation notable. Ces caractéristiques sont présentées en table 9.9.

La seconde visualisation, à droite en figure 9.33, représente les liens d'amitié autour de la même communauté. Les liens y sont tous réciproques ; l'un des membres du groupe se retrouve isolé, en

9. Il revendique être en train de construire un site pornographique : « building a porn site cuz i can ».

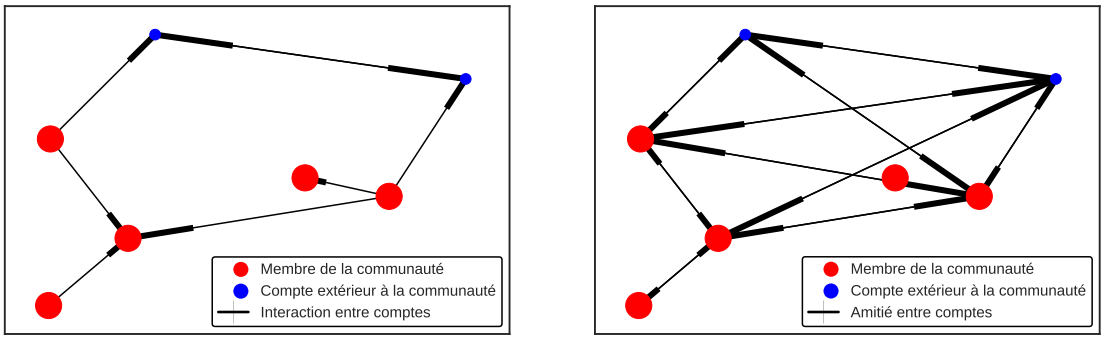



FIGURE 9.33 – Une petite communauté et sa frontière, à gauche dans le graphe des interactions $G_I$ ; à droite dans le graphe des amitiés $G_F$

TABLEAU 9.9 – Caractéristiques du groupe « Bishop » sur $G_I$

| Mesures topologiques | | | | Mesures thématiques | | |
|---|---|---|---|---|---|---|
| Taille | d | c | TPR | ξ | ρ | $\theta f.igf$ |
| 5 | 0.4 | 0.2 | 0.0 | 0.4 | 0.017 | 0.47 |

haut à droite : sa présence dans la communauté est purement fortuite. À l'opposé, les deux nœuds considérés comme extérieurs lors de l'analyse de $G_I$ pourraient être intégrés au groupe au vu des liens d'amitié ; il faut cependant prendre en compte leur environnement auparavant.

#### 9.3.5.4 Graphe des objets sociaux

Le troisième graphe exploité durant cette étude de cas est construit à partir des échanges d'objets sociaux, ici les images, sondages et pages de discussion. Ce graphe, noté $G_\Omega$, est propice à l'apparition de communautés, car il repose sur l'action conjointe de comptes sur les mêmes supports (par exemple, une page de sondage, qui suscite des votes et des commentaires).

L'analyse de ces communautés par les mesures thématiques montre, en figure 9.34, une certaine faiblesse de cohésion thématique. Aucun groupe ne dépasse $\xi = 0.25$, et les valeurs de ρ, en échelle logarithmique, sont basses. Ces scores s'expliquent par la déconnexion entre les thématiques détectées sur les publications textuelles (billets sur *The Wire* ou sur les pages personnelles), qui servent à calculer ces mesures thématiques, et les objets sociaux qui relient ces comptes entre eux.

L'étude des autres mesures vient compléter cette vue. La figure 9.35 représente les valeurs de pertinence thématique $\theta f.igf$ en abscisses, contre la cohésion structurelle TPR en ordonnées. La majorité des communautés sont cantonnées dans le coin inférieur gauche, ce qui s'interprète par la faiblesse structurelle et thématique de ces groupes. Cependant, l'une des communautés, le groupe n°9, se distingue du reste par un TPR = 0.3 et $\theta f.igf = 0.51$, le plaçant dans le coin supérieur droit. Ces scores ne sont pas particulièrement élevés, mais suffisent à démarquer cette communauté.

Une visualisation de ce groupe est proposée en figure 9.36, où les nœuds sont colorés en vert, car ils ont tous adopté le $role_1$, décrit en section 9.3.4.3. En effet, ce groupe n'abrite pas d'utilisateur « hyper-émetteur » (rôle 0). Dans le détail, l'exploration des contenus publiés et partagés dans ce groupe met en avant deux pages de discussion hispanophones, *que-buscas-en-la-deep* (« que cherches-tu dans le Deep Web ») et *ventajas-de-tor* (« les avantages de TOR »).

Cette section a mis en évidence la capacité du système à construire des graphes à partir de relations fournies par la plate-forme de manière explicite (les liens d'amitié, la création d'objets sociaux) ou implicite (les mentions « cachées » dans les messages). Nous avons exploré ces graphes au travers

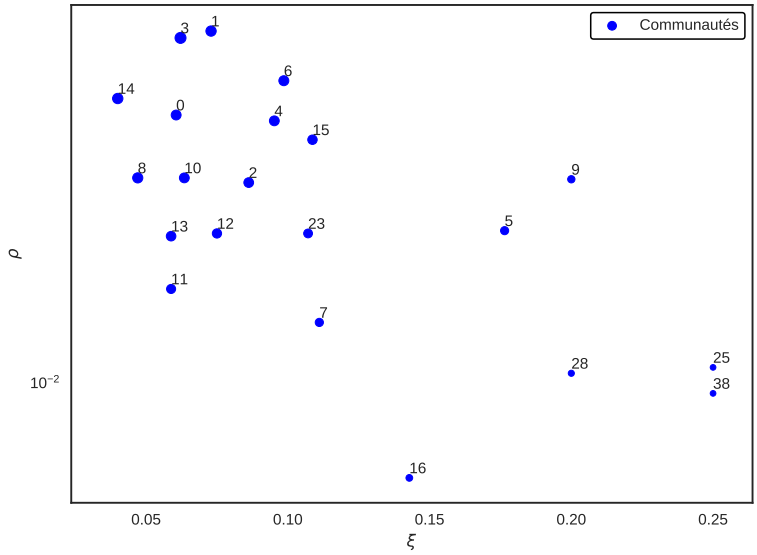



FIGURE 9.34 – Comparaison de la cohésion thématique ξ et de la représentativité ρ des communautés issues de $G_\Omega$

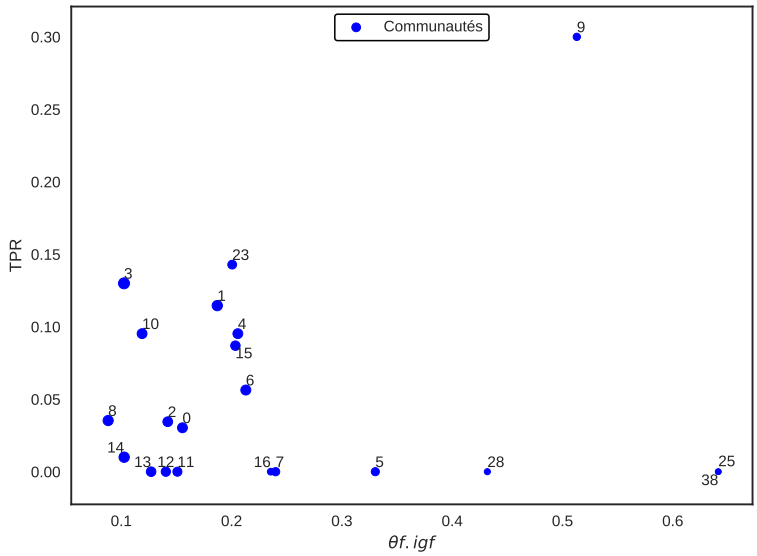



FIGURE 9.35 – Comparaison de la pertinence thématique $\theta f.igf$ et de la cohésion structurelle TPR des communautés issues de $G_\Omega$

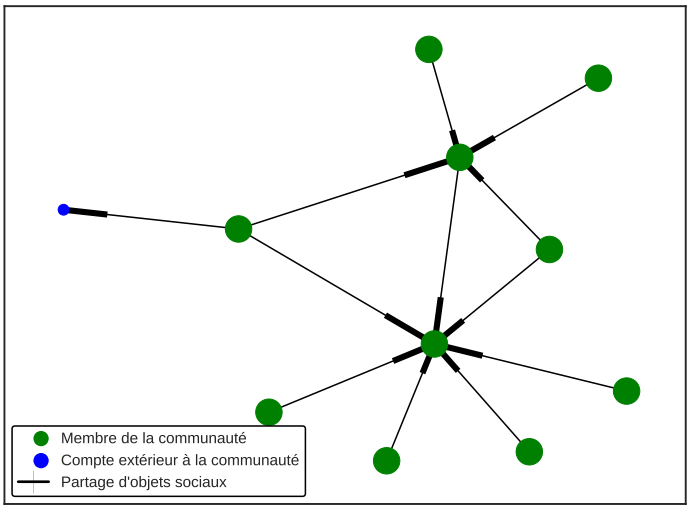



FIGURE 9.36 – Structure et rôles-types dans le groupe 9 issu de $G_\Omega$

des communautés qui y sont détectées : les scores topologiques et thématiques quantifient la cohésion structurelle et sémantique des groupes, dont une visualisation est disponible.

### 9.3.6 Conclusion de l'application à Galaxy2

L'analyse des réseaux sociaux numériques se focalise sur les plate-formes les plus connues, parmi lesquelles Facebook et Twitter. Cependant d'autres instances existent, parfois bien cachées : Galaxy2 est bien plus petit que Twitter, mais suffisamment grand pour rendre difficile son analyse « à la main », via un navigateur web. Sur presque trois ans (janvier 2015 - septembre 2017), ce sont presque 20 000 comptes créés, et plus de 30 000 publications.

Cette quantité de données est exploitée par le système SARTN en trois phases. La détection de thématiques et de sentiment fournissent des informations pratiques pour évaluer la teneur des 30 000 contenus textuels émis. À propos des comptes utilisateurs, la situation est particulière. La spécificité de Galaxy2, par son aspect confidentiel, en fait un étrange média participatif : comme les gens n'y parlent pas de leur vie quotidienne, l'ensemble des comptes vraiment actifs s'en trouve réduit. L'étude comportementale révèle pourtant des façons différentes d'agir et d'influer. Enfin, les différentes fonctionnalités proposées par la plate-forme résultent en différentes vues sous forme de graphes, permettant d'analyser le réseau par les communautés qui y émergent, puis de qualifier ces communautés par leurs caractéristiques topologiques, thématiques, par les rôles-types de leurs membres, et par les principaux contenus qui y sont publiés et échangés. Les fonctionnalités différentes imposent de modifier la façon de construire les graphes, et éventuellement de les combiner. Alors qu'il est intuitif de se baser sur les interactions et mentions dans le cas de Twitter, ce support est moins pertinent sur Galaxy2.

Cette étude de cas nous a permis de montrer que les fonctionnalités de SARTN s'appliquent aussi efficacement sur d'autres réseaux sociaux numériques. Quelques adaptations sont certes nécessaires, les données n'étant pas nommées de la même façon ; toutefois ni l'approche ni les algorithmes n'ont été modifiés. Nous estimons donc possible et intéressant d'analyser d'autres plates-formes, car le comportement des utilisateurs et la logique d'influence n'y apparaissent pas toujours de la même manière.

## 9.4 Synthèse de l'évaluation

Pour récapituler les forces et limites du système SARTN, nous faisons référence aux critères définis en section 9.1, qui ont permis de structurer les deux expériences, sur Twitter et sur Galaxy2. Ces deux jeux de données sont comparables en structure : il s'agit de messages et actions publiées sur des plates-formes de micro-blogging, dont les textes sont majoritairement courts et écrits en anglais. Une comparaison des tailles des corpus et de leur variété est proposée en table 9.10. Toutefois, il faut souligner la grande variété des types d'actions disponibles sur Galaxy2, qui multiplie les types de liens entre utilisateurs. Les jeux de données, bien que trop grands pour être explorés « à la main », ne relèvent pas au sens propre du *Big Data* et ne requièrent pas des ressources distribuées pour être traités.

TABLEAU 9.10 – Comparaison des deux études de cas

| | Utilisateurs | Textes | Durée | Type d'éléments | Graphes |
|---|---|---|---|---|---|
| *KevRandTweets* | 734 888 | 9 671 711 | 1 mois | Tweets | $G_I$, $G_\Omega$, $G_\Theta$ |
| *Galaxy2* | 19 233 | 29 816 | 3 ans | Toutes actions | $G_F$, $G_I$, $G_\Omega$ |

Le critère 1, orienté sur l'exploitation des contenus textuels, requiert l'**extraction des opinions** des messages, c'est-à-dire leur cible et leur polarité. Celle-ci est obtenue par la détection du sentiment ; la thématique représente une approximation de la cible de l'opinion du message, et permet de limiter le nombre d'entités concernées, au détriment de la précision. Malgré les outils d'exploitation des textes fournis, ce critère n'est donc que partiellement atteint. Des pistes sont identifiées pour mieux exploiter ces ressources textuelles, parmi lesquelles l'étude de la temporalité des mots-clés, ainsi que la catégorisation des URLs et des images partagées.

Tant sur Twitter que sur Galaxy2, les différents **indicateurs d'influence** construisent une vue complète des acteurs-clés du réseau, que ce soit en termes d'actions réalisées, de popularité ou de position de référence dans le graphe des mentions. Ces scores facilitent l'exploration des jeux de données, et permettent dans un second temps de comparer les influences respectives de deux comptes, satisfaisant le critère 2. La différence d'échelle entre les deux réseaux sociaux rend évidente la pertinence du paramétrage des scores de popularité et d'expertise.

Les **comportements** sont efficacement regroupés et agrégés en profil-type pour les jeux de données, fournissant à l'utilisateur de SARTN les outils de catégorisation. Si les mêmes tendances se dégagent, avec notamment une concentration des publications par une minorité d'acteurs, il faut noter que la répartition est différente, Galaxy2 ayant plus de difficultés à susciter l'adhésion et la participation des utilisateurs. En accord avec le critère 3, l'information comportementale se révèle différente du rôle, calculé sur le graphe des interactions, qui facilite la lecture de la structure interne des communautés. Nous pensons toutefois que la catégorisation des comportements en profils-types peut encore gagner en lisibilité et légitimité, par exemple en la déclinant pour chaque aspect comportemental : plutôt que de combiner l'ensemble des caractéristiques proposées, l'étude des comptes sur l'un des aspects résulterait en des étiquettes plus faciles à interpréter.

Sur les deux jeux de données, des exemples de construction, d'exploitation et de combinaison des **graphes** sont proposés. S'ils contiennent toujours autant de nœuds (les comptes présents dans le corpus), ces graphes présentent toutefois des structures différentes, selon leur nature : les graphes d'interaction, d'abonnement, de partage présentent des connectivités variées et résultent en des communautés détectées très différentes. Ainsi, c'est l'utilisateur de SARTN lui-même qui vérifie le critère 4 : il n'y a pas de solution unique aux applications bénéficiant de l'analyse des graphes sociaux, si ce n'est le recours aux outils de construction puis de filtrage des arcs et nœuds, pour tirer profit des informations à disposition.

L'axe majeur de l'exploitation des graphes consiste en la **détection de communautés**, réalisée

selon l'algorithme Louvain. Par rapport au jeu de données *KevRandTweets*, le corpus Galaxy2 est plus petit, mais aussi étalé sur une période temporelle beaucoup plus large, presque trois ans (janvier 2015 - septembre 2017), ce qui met l'accent sur la temporalité. Bien que SARTN conserve les informations temporelles, il n'en tire pas profit : les graphes sont construits en mélangeant des interactions, pour $G_I$, par les échanges autour d'objets sociaux pour $G_\Omega$, ou autour des amitiés pour $G_F$, mais ces informations sont parfois très éloignées dans le temps. Nous pensons que cette dimension temporelle devrait être prise en compte lors de la détection des communautés lors de travaux futurs.

Ces groupes sont ensuite caractérisés par des mesures topologiques et thématiques. Il est possible de voir comment ces communautés interagissent sur le réseau social, et de visualiser l'une de ces communautés, éventuellement en enrichissant cette vue par les rôles de ses membres. Si cela permet de considérer le critère 5 comme satisfait, il reste cependant des axes de recherche : la détection de communautés devrait inclure une temporalité des communautés, et éventuellement permettre l'affiliation d'un nœud à plusieurs groupes.

Enfin, la caractérisation des communautés grâce aux **mesures de cohésion thématique** introduites en chapitre 7 permet d'estimer l'influence des groupes, en termes de nombre d'utilisateurs intéressés, ou de quantité de contenus émis, sur une thématique. Ces mesures enrichissent la vue déjà proposée par les mesures purement topologiques, issues de l'analyse des graphes, répondant au critère 6. Pourtant nous pensons qu'un « score d'influence » global, au niveau du groupe, présente un intérêt certain ; il pourrait combiner ces informations thématiques et topologiques, en lissant l'impact de la taille du groupe.

Pour conclure ce chapitre, le système SARTN répond aux critères exposés mais présente des axes d'amélioration possible et nécessaire : en ingénierie, par l'intégration de modules existants pour réaliser des traitements, mais aussi du point de vue scientifique, pour affiner la compréhension du texte, l'étude des comportements, et la caractérisation des communautés sociales présentes sur le réseau.

# Quatrième partie

# Conclusion et perspectives

# CHAPITRE 10

# Conclusion et perspectives

Ce chapitre clôture ce document : une première section rappelle nos contributions théoriques à l'analyse des réseaux sociaux numériques, dans la détection de l'opinion, l'étude des comportements des comptes utilisateurs, et la caractérisation de communautés par des mesures de cohésion thématique. La section 10.2 discute des pistes d'amélioration de ces contributions, ainsi que des limites du système implémenté. Finalement, la section 10.3 propose des perspectives de travaux de recherche futurs.

## 10.1 Synthèse sur les contributions

Les réseaux sociaux numériques sont déjà bien installés dans nos quotidiens ; pourtant nous sommes tous encore démunis pour les appréhender dans leur ensemble. La quantité de messages émis rend nécessaires des outils visant à extraire de ces messages leur essence : les opinions qu'ils contiennent. De même, la profusion d'émetteurs de contenus s'accompagne du besoin d'outils pour identifier et comparer les comptes, afin de découvrir efficacement les comptes importants, et les thématiques sur lesquelles ils sont reconnus. Enfin, les réseaux sociaux sont aussi des systèmes complexes, propices à l'émergence de groupes plus ou moins coordonnés.

Pour répondre à ces défis et difficultés, nous avons énoncé ces besoins d'outils de compréhension de l'espace médiatique par notre problématique « détection des opinions, acteurs-clés et communautés thématiques dans les médias sociaux ». Ainsi, les besoins sont répartis sur trois niveaux d'analyse, chaque niveau alimentant les autres : nous étudions les messages, les comptes, et les groupes de comptes.

La détection d'**opinions** consiste à détecter la *polarité* du message concernant une *cible*. Bien que nous considérions les *thématiques* en chapitre 9, qui sont plus générales et englobent l'ensemble des cibles mentionnées, nous avons exploré une autre piste, avec la détection de posture, dans le cadre de notre participation au défi SemEval 2016. Afin de mieux tirer profit du corpus d'apprentissage fourni, nous avons proposé de ré-exploiter et d'adapter aux tweets une méthode de désambiguïsation sémantique : les contextosets. Sans supervision, cette approche construit une ressource linguistique semblable aux *synsets*, qui est rendue nécessaire par l'usage et l'abus de néologismes, hashtags et orthographes inventives. Cette ressource est ensuite exploitée pour clarifier les textes, et en déterminer la posture par apprentissage automatique.

L'analyse des **comptes** utilisateurs englobe plusieurs besoins et applications. Dans un premier temps, l'identification d'acteurs-clés dans un jeu de données requiert une mesure de l'influence, qui

prend plusieurs formes. Nous l'avons décomposée en popularité, influence et expertise, couvrant trois notions distinctes mais souvent confondues. Dans un second temps, l'étude des comportements reconnaît les différents niveaux et types d'activité des comptes, dont l'usage fait des hashtags, de la publication d'images et du partage de liens Web ; ces informations sont stockées dans des profils qui sont ensuite agrégés en profils-types, facilitant la reconnaissance de comportements similaires. À ces données numériques s'ajoute l'exploitation du graphe social, par l'établissement d'une typologie des positions : les rôles-types. La combinaison de ces indicateurs permet à un analyste de quantifier et comparer facilement les utilisateurs, et d'identifier les comptes d'intérêt parmi la population d'émetteurs de publications.

L'apparition de **groupes** ou communautés est naturelle dans tous les réseaux sociaux, qu'ils soient numériques ou non. De nombreuses définitions co-existent, répondant à des approches et applications différentes ; dans nos travaux, les communautés sont définies à partir des interactions entre les comptes utilisateurs. Là où de nombreux travaux proposent des méthodes de détection, nous nous focalisons sur la description des groupes de comptes. Nous avons ainsi proposé des mesures pour évaluer la cohésion thématique de ces groupes ; ces mesures viennent naturellement compléter les mesures topologiques existantes, quantifiant la force de liaison interne du groupe. De cette manière, l'ensemble du réseau est décrit par l'activité sociale qui s'y déroule, enrichie par les indicateurs de cohésion thématique et structurelle.

Ces contributions sont implémentées dans un système, SARTN. Couvrant trois niveaux d'analyse, sur le plan des textes, des comptes utilisateurs et des communautés, il répond aux critères fonctionnels définis en chapitre 9. Cet outil logiciel permet d'agréger les thématiques et sentiments des messages, de caractériser le comportement et d'évaluer l'influence des comptes, et il mesure la cohésion des communautés, permettant d'en qualifier quelques-unes de « thématiques ».

Pour mettre en évidence les forces et limites de ces contributions, nous avons réalisé deux études de cas, utilisant le système SARTN, et portant sur deux réseaux sociaux numériques : Twitter, et Galaxy2. Le premier jeu de données est composé de tweets ; il ne contient qu'une infime fraction des contenus publiés sur Twitter, mais correspond aux limites techniques de la collecte autorisée par la plate-forme. Le second corpus est composé de l'intégralité des activités publiées sur Galaxy2, un réseau social hébergé sur TOR, mais est de taille bien plus modeste. Ces deux études de cas utilisent tous les outils déployés au niveau de l'analyse des textes, de l'étude des comptes utilisateurs et de la caractérisation des communautés thématiques, et permettent d'évaluer la pertinence et le fonctionnement de nos contributions.

## 10.2 Pistes d'amélioration des contributions

Nos contributions ne sont pour autant pas parfaites ; nous en rappelons ici les limites, niveau par niveau, auxquelles s'ajoutent des limites plus globales, touchant l'implémentation du système et son intégration dans les outils existants.

La détection d'**opinions** produit encore des scores peu fiables, dépendants des corpus d'apprentissage : si la performance des classifieurs de revues de produits ou de films s'est grandement améliorée ces dernières années, il n'en va pas de même pour mesurer l'opinion de tweets, c'est-à-dire de messages qui ne sont pas rédigés dans l'objectif de partager expressément cette opinion (au contraire des revues de produits). *A fortiori*, la détection de posture politique dans des messages courts présente plusieurs difficultés, dont la spécificité du vocabulaire, et son évolution rapide au long d'une campagne électorale. Certes, notre modèle de désambiguïsation par les contextosets améliore de quelques points la qualité d'un classifieur statistique, mais il ne permet finalement que de signaler le défi de l'extraction de l'opinion. Notons que ce domaine est en pleine effervescence, et des progrès substantiels sont espérés dans les prochaines années.

Les opinions détectées sont émises par des utilisateurs qui bénéficient (ou pâtissent) de visibilités différentes. La littérature propose déjà de nombreux **scores d'influence**, auxquels nous ajoutons trois indicateurs : popularité, influence et expertise. Tant pour l'état de l'art que pour notre proposition, la validation de ce module est ardue : les listes d'influenceurs sont toutes sujettes à caution et ne représentent qu'une facette de l'influence. Plutôt qu'un indicateur-miracle, notre contribution consiste ici en l'exploitation conjointe de trois valeurs, répondant à trois points de vue sur la notion « d'acteur-clé ».

La caractérisation des comportements par les **profils-types** et des positions sociales des comptes par les **rôles-types** couvrent deux aspects complexes des réseaux sociaux. Cependant, notre contribution résulte en des étiquettes de profils et rôles *ad hoc* issues d'un apprentissage non-supervisé : ils diffèrent d'un jeu de données à l'autre. Cette adaptabilité rend nécessaire un module facilitant les nouvelles interprétations des profils-types et rôles-types par l'utilisateur de SARTN.

La détection et caractérisation des **communautés** thématiques présente également quelques limites. Pour la détection tout d'abord, le choix de l'algorithme Louvain, qui répond aux exigences en termes de modularité finale du partitionnement et de temps de calcul, soulève des difficultés. Certains groupes contenant plusieurs milliers de comptes pourraient encore être scindés. La temporalité des interactions n'est pas prise en compte par cette méthode ; enfin les comptes ne peuvent appartenir qu'à un unique groupe. Dans un second temps, la phase de caractérisation montre une sensibilité à la taille des groupes : un indicateur seul n'est pas exploitable, il faut absolument en combiner plusieurs. Finalement, une dernière limite : si la cohésion thématique est bien mesurée, la dispersion des polarités de sentiment ne l'est pas. Au-delà d'une cohésion thématique, il serait intéressant de mesurer une cohésion d'opinions, et son évolution temporelle.

D'un point de vue industriel, l'intégration de nos contributions en un système, SARTN, nécessite des développements supplémentaires : si les jeux de données traités sont relativement grands, ils ne sont toutefois pas à l'échelle du *Big Data*. De plus, il y manque des outils pour répondre aux besoins d'analyse, parmi lesquels l'étude de l'historique d'un hashtag ou d'un compte, le rapprochement de comptes entre plusieurs réseaux sociaux, l'exploration des sites Web, l'analyse des images reçues, ainsi qu'un outil de collecte intelligente (par opposition à la liste figée de 5000 comptes utilisée pour *KevRandTweets*).

Ces travaux de thèse n'ont pas vocation à devenir un produit commercialisé, mais bien à s'intégrer dans la solution développée par AIRBUS, et nommée Fortion®MediaMining . La chaîne de traitement proposée doit donc être validée et figée, une fois que son ergonomie et sa façon de s'interfacer avec les autres modules seront déterminées.

## 10.3 Travaux futurs et axes d'ouverture

Nous distinguons deux niveaux de perspectives pour réaliser l'avenir de cette thèse. Le premier niveau concerne directement les fonctionnalités proposées par SARTN, afin notamment d'accompagner son utilisation par les utilisateurs. Le second niveau se base sur trois grands axes de réflexion à plus long terme.

### 10.3.1 Perspectives d'utilisation des fonctionnalités de SARTN

La mise à disposition de nos contributions dans un système unique ouvre des pistes d'utilisation applicative, décrites en figure 10.1. Ainsi, le système exposé sert de socle à ces fonctionnalités supplémentaires.

L'utilisation des informations extraites (objets sociaux, méta-données autour du texte, interaction, profils des utilisateurs) rend possible le calcul de *similarité* selon l'aspect d'intérêt dans le cas

d'usage : la *catégorisation* des utilisateurs, et l'*extension de requêtes* (vers des thématiques, comptes ou groupes similaires), sont des briques qui viendront naturellement donner de la clarté dans l'exploration des réseaux sociaux. Une fois trouvé un acteur-clé d'intérêt, il devient aisé de trouver des comptes ayant un comportement ou des données biographiques semblables.

La donnée disponible (qu'elle soit brute ou calculée) peut être chargée sous forme de série temporelle, voire de série géo-temporelle [1], ce qui ouvre la voie à la *détection de comportement anormal* dans le temps.

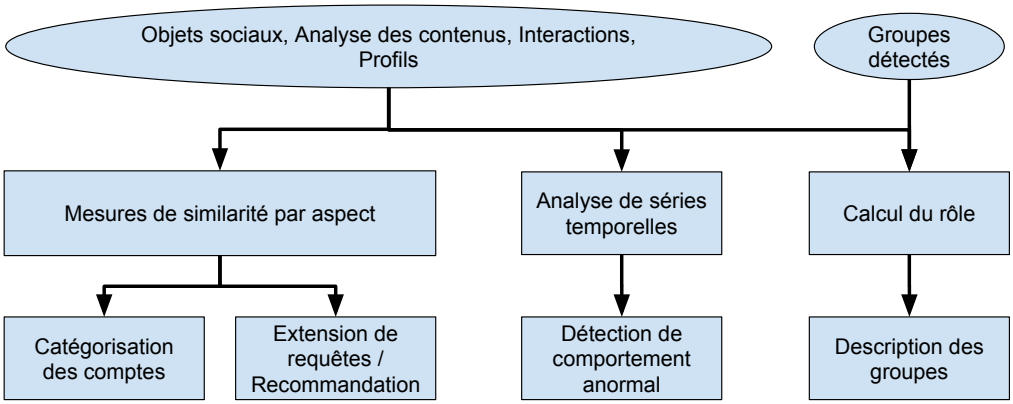



FIGURE 10.1 – Perspectives d'ouverture des fonctionnalités

Enfin, l'association des données comportementales avec les groupes détectés permet de calculer des rôles des utilisateurs dans le groupe, et de *décrire les groupes* par la présence ou par la répartition de type d'acteurs au sein du groupe. Certains groupes sont centrés sur un meneur, d'autres suivent des approches beaucoup plus distribuées et collaboratives.

Outre ces ajouts, nous pensons pertinent de concevoir une nouvelle approche de détection des communautés thématiques. Des travaux futurs porteront sur les algorithmes d'obtention de couvertures par l'optimisation des mesures de cohésion structurelle (*TPR*) et thématique ($\theta f.igf$). Nous pensons ainsi obtenir un plus grand nombre de communautés, elles-mêmes plus pertinentes.

Ces extensions et améliorations des fonctionnalités déjà implémentées faciliteront l'utilisation du système, et donneront vraisemblablement lieu à de nouvelles pistes de recherche pour comprendre et exploiter des notions rendues plus faciles d'accès, dont le rôle dans le groupe. En plus de ces perspectives fonctionnelles, il nous faut ajouter quelques mots sur trois grands axes de recherche future.

### 10.3.2 Axes d'ouverture

Peu évoquée dans ce document, l'**éthique** constitue le premier axe d'ouverture : bien que la donnée issue du Web social soit aisément accessible, elle est pourtant souvent personnelle, et en conséquence elle est protégée en France par la Commission Nationale de l'Informatique et des Libertés depuis 1978, et par le Règlement Général de Protection des Données, adopté en Union Européenne en 2016. La question consiste alors à trouver le moyen éthique et légal de tirer parti des données disponibles, en respectant la liberté de chacun. La perception de cette liberté évolue dans le temps, et de nombreux traitements automatiques, souhaités il y a quelques années, sont désormais tabous ou interdits.

1. sur certaines plates-formes, il est courant de voir une position GPS associée à un message

Le deuxième axe d'ouverture consiste en la gestion de la **temporalité**, grande absente de nos travaux. Nous avons la conviction qu'il y a beaucoup à apprendre à étudier l'évolution du langage, des comportements, des actions d'influence, et des groupes de comptes dans le temps. La visualisation des indicateurs calculés dans le temps est certes aisée à réaliser à court terme ; il n'en va pas de même, par exemple, de l'attribution de l'étiquette « d'influenceur » en fonction des actions temporelles réalisées, qui nécessite un modèle spécifique.

Troisième perspective, l'analyse de l'objet « média social » doit encore se poursuivre sur des plate-formes et **modes de communication** encore relativement peu explorés : d'une part, les agrégateurs dont *Reddit* nécessitent une étude conjointe des contenus publiés et des commentaires afin de trouver les informations recherchées et de qualifier le comportement des comptes. D'autre part, les plate-formes de discussion ou de chat, dont *Telegram*, *Whatsapp* ou *Discord* nécessitent des modèles d'étude de la discussion, et de mesure des relations d'opposition ou de confiance.

Ces perspectives seront explorées, notamment dans le cadre de l'axe de recherche autour des réseaux sociaux numériques, développé par l'équipe MIND au LITIS, mais aussi en entreprise, comme chez AIRBUS dans le cadre de thèses, de projets internes ou encore collaboratifs. Nous espérons qu'ainsi augmentera notre compréhension de l'espace informationnel, et des mécanismes d'interaction entre membres de réseaux sociaux.


Roman von Ungern-Sternberg
*« En avant, à la recherche de nos folies et de nos gloires ! »*

# Cinquième partie

# Annexes

# ANNEXE A

## Liste des publications

— Extracting contextonyms from Twitter for stance detection, *G. Gadek, J. Betsholtz, A. Pauchet, S. Brunessaux, N. Malandain and L. Vercouter*, ICAART, 2017, Volume 2, 132-141.

— Topical cohesion of communities on Twitter, *G. Gadek, A. Pauchet, N. Malandain, K. Khelif, L. Vercouter and S. Brunessaux*, KES, 2017, 10p. **Best student paper award.**

— Measures for topical cohesion of user communities on Twitter, *G. Gadek, A. Pauchet, N. Malandain, K. Khelif, L. Vercouter and S. Brunessaux*, Web Intelligence, 2017, p211-218.

— AI techniques to analyse a social network on text, user and group level : application on Galaxy2, *G. Gadek, A. Pauchet, S. Brunessaux, K. Khelif and B. Grilheres*, APIA, 2018, 9p.

— Topological and topical characterisation of Twitter user communities, *G. Gadek, A. Pauchet, N. Malandain, L. Vercouter, K. Khelif, S. Brunessaux and B. Grilheres*, Data Technologies & Applications, 2018, 20p.

— Application of AI Techniques to Deep Web Social Network Analysis, *G. Gadek, S. Brunessaux and A. Pauchet*, NATO Specialists' Meeting - IST 160, 2018, 16p.

# ANNEXE B

## Annexes techniques à propos des implémentations

### B.1 Mots-clés pour le corpus *GenTweets*

Pour recueillir les tweets composant le corpus *GenTweets*, nous avons ouvert une requête sur l'API Stream de Twitter contenant les mots suivants :

— Atheism :
god, hope, lord, God, halal, haram, Faith, faith, atheism, Atheism, bible, Bible, quran, Quran, religion, Religion, religions, Religions, #freethinker, Jesus, temple, Temple

— Climate :
climate, Climate, climate change, carbon, glacier, global warming, CO2, COP21, sea level

— Feminism :
feminists, Feminists, feminism, women's rights, gender, gender equality, sexism, sexist

— Hillary :
woman president, Hillary, clinton, Clinton, #WhyImNotVotingForHillary, #NoHillary2016, #StopHillaryClinton2016, HillaryClinton

— Abortion :
abortion, pro-choice, pro-life, #ProLifeYouth, #AllLivesMatter, unborn, pregnant, #womenshealth, #womensrights

## B.2 Données complètes de l'exemple *ArtsTweets*

Cette annexe contient les messages créés pour illustrer la contribution du chapitre 7. La Table B.1 contient les messages bruts : identifiants des messages et de leurs auteurs, et texte du message. La Table B.2 contient les informations extraites de ces messages : interactions explicites, partages d'objets sociaux, et thématiques.

TABLEAU B.1 – Ensemble de messages pour l'exemple illustratif *ArtsTweets*

| Message $ID_t$ | Auteur $ID_u$ | Contenu du message |
|---|---|---|
| 101 | 1 | hello @2 viens voir des films sur mon site, le meilleur du ciné : http ://url1 |
| 102 | 1 | hello @3 viens voir des films sur mon site, le meilleur du ciné : http ://url1 |
| 103 | 1 | hello @4 viens voir des films sur mon site, le meilleur du ciné : http ://url1 |
| 104 | 1 | hello @6 viens voir des films sur mon site, le meilleur du ciné : http ://url1 |
| 105 | 1 | hello @7 viens voir des films sur mon site, le meilleur du ciné : http ://url1 |
| 106 | 2 | Rien de tel qu'une sortie au cinéma ! |
| 107 | 3 | arrête ton spam @1 ! tes films sont nuls pas en HD |
| 108 | 3 | j'attends Godot, et vous ? à la scène nationale 76 ! |
| 109 | 4 | la prof de français et son theaaaaaatre xD la flemme quoi |
| 110 | 5 | site rustique mais pratique http ://url1 |
| 111 | 5 | jazzy mais pas groovy, je kiffe la CompilJazz ! |
| 112 | 6 | montage de quintett de jazz 101 : trouver un sax |
| 113 | 7 | on est en 2018 et y'en a qui écoutent encore du jazz ! et pk pas du Trénet tsss |
| 114 | 8 | le site de @1 est pas mal pour éviter de rakker au ciné |
| 115 | 8 | hey les potos on se fait un cinoche ? #ArseneLupinLeFilm |
| 116 | 9 | @12 viens voir #ArseneLupinLeFilm samedi 20h |
| 117 | 9 | le théâtre c'est bien aussi |
| 118 | 9 | Molière au Grand Théâtre ! j'ai hâte ! |
| 119 | 10 | @15 la bande-son de #ArseneLupinLeFilm : ça c'est dla musique jazz ! |
| 120 | 11 | #ArseneLupinLeFilm le super navet |
| 121 | 11 | mon frangin qui écoute du slipknot il est trop dark |
| 122 | 11 | j'aime quand ils passent du rammstein à la radio |
| 123 | 13 | @12 viens avec moi au #Hellfest |
| 124 | 14 | t'es dispo le week-end du 21 ? y'a #Hellfest @12 |
| 125 | 13 | @14 tu y vas 2 ou 3j ? #métal |
| 126 | 15 | je t'aime @12 ! allons au #Hellfest ensemble ! |
| 127 | 16 | super programmation au #Hellfest c't'année |
| 128 | 16 | c'est de la drague ou de l'humour @15 ? |
| 129 | 12 | je préfère regarder un bon film |
| 130 | 17 | L'affiche de StarShip Trooper elle claque ! affiche.png |
| 131 | 18 | @17 je plussoie elle déchire affiche.png |
| 132 | 18 | Racine vs Molière, le match de l'année |
| 133 | 18 | @19 tu viens voir la pièce où joue ma fille ? c'est jeudi soir |
| 134 | 19 | répèt' pour notre concert à jazz sous les pommiers #fausseNote |

TABLEAU B.2 – Ensemble d'informations extraites des tweets dans *ArtsTweets*

| Message ID$_t$ | Auteur ID$_u$ | Partages ω | Interactions $i$ | Thématiques θ |
|---|---|---|---|---|
| 101 | 1 | url1 | ME → @2 | ciné |
| 102 | 1 | url1 | ME → @3 | ciné |
| 103 | 1 | url1 | ME → @4 | ciné |
| 104 | 1 | url1 | ME → @6 | ciné |
| 105 | 1 | url1 | ME → @7 | ciné |
| 106 | 2 | | | ciné |
| 107 | 3 | | ME → @1 | ciné |
| 108 | 3 | | | théâtre |
| 109 | 4 | | | théâtre |
| 110 | 5 | url1 | | ciné |
| 111 | 5 | | | jazz |
| 112 | 6 | | | jazz |
| 113 | 7 | | | jazz |
| 114 | 8 | | ME → @1 | ciné |
| 115 | 8 | #ArseneLupinLeFilm | | ciné |
| 116 | 9 | #ArseneLupinLeFilm | ME → @12 | ciné |
| 117 | 9 | | | théâtre |
| 118 | 9 | | | théâtre |
| 119 | 10 | #ArseneLupinLeFilm | ME → @15 | jazz |
| 120 | 11 | #ArseneLupinLeFilm | | ciné |
| 121 | 11 | | | métal |
| 122 | 11 | | | métal |
| 123 | 13 | #Hellfest | ME → @12 | métal |
| 124 | 14 | #Hellfest | ME → @12 | métal |
| 125 | 13 | | ME → @14 | métal |
| 126 | 15 | #Hellfest | ME → @12 | métal |
| 127 | 16 | #Hellfest | | métal |
| 128 | 16 | | ME → @15 | métal |
| 129 | 12 | | | ciné |
| 130 | 17 | affiche.png | | ciné |
| 131 | 18 | affiche.png | | ciné |
| 132 | 18 | | | théâtre |
| 133 | 18 | | ME → @19 | théâtre |
| 134 | 19 | #fausseNote | | jazz |

## B.3 Galaxy2 : les raisons de la rupture

Nous proposons ici le message que *Lameth*, fondateur du réseau social Galaxy2, a laissé à l'adresse du site [1] depuis que le service est tombé à l'automne 2017.

So, it finally happened. The server broke down and your terrible host here (me, not the current host, mind you !) hadn't been keeping regular backups off the server... I'm trying to see what I can salvage, but to be honest with you guys, then I'm not very optimistic.

The admins and a few other users has for a long time now been privy to my plans about either handing over G2 for someone else to run, or shutting down G2 completely. Seems like the server got tired of waiting for me to get my shit together and took the decision for me.

Why ? Well, a couple of reasons, really, but the major one being purely selfish. I knew I was doing a shit job hosting G2, and as it grew more and more popular, so did my guilt and stress about not being a proper host and a proper admin. I couldn't dedicate the time for running and managing G2 that I felt it deserved. So at some point I came to a conclusion ; I think the mature decision was to give it up and hand it over to someone more capable, for the sake of the community. Turns out I also have a terrible habit of never actually doing the stuff I intend to do, so even the process of handing over the reigns to another host never really got further than me putting out a few feelers and questions to a few people.

So now we're here. G2 seems to be done, although I try and find time to see if I can manage to salvage something that might make the admins and the new host capable of continuing G2.

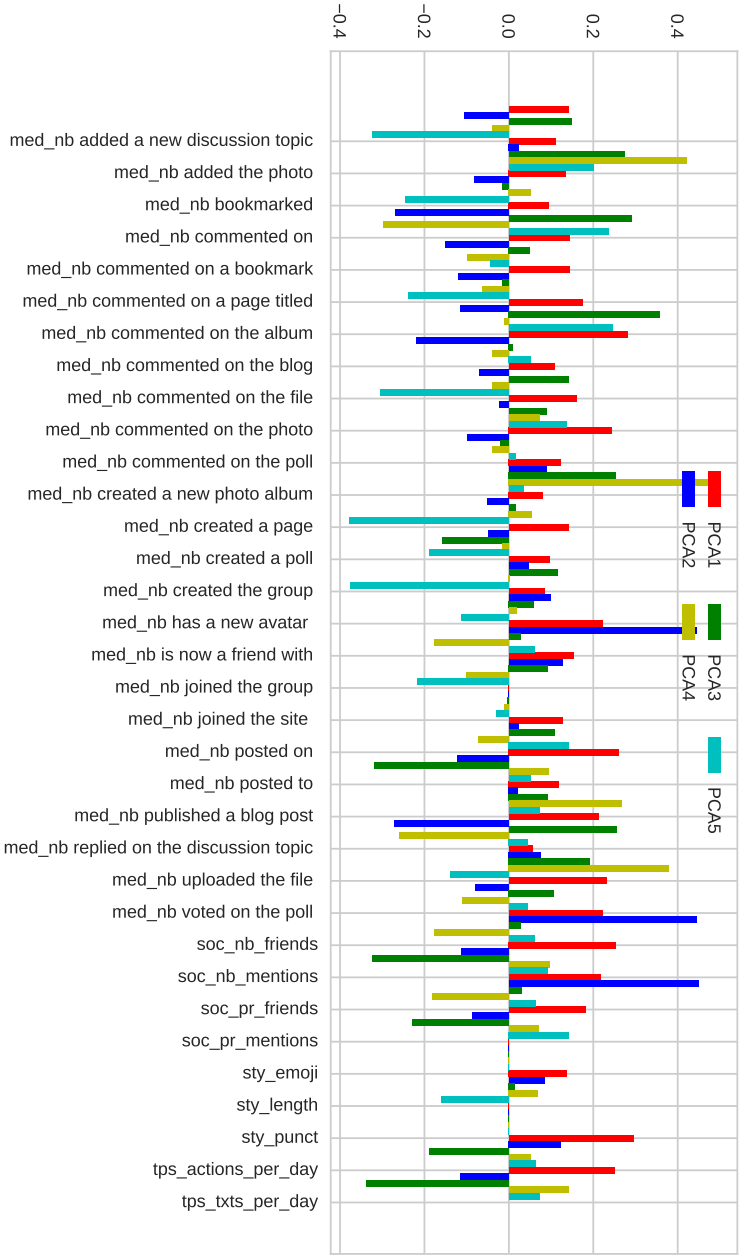
I'd like to encourage people to migrate out to other Tor Hidden social Services, regardless of whether G2 can be recovered or not. I hope other social sites will pop up (Galaxy3, anyone ?), but I believe there are still some of the "older" chat services around. They might be a good place to reconnect with other G2 users.

Even if G2 is recovered and continues, then this is the end of the road for me. I have little time to spare for this or any online community due to family and work stuff. It's a bittersweet farewell for me. One one hand, it lifts a burden from my shoulders that I don't care to carry around any longer ; on the other hand I really did enjoy being a part of this community, of being part of the beginning and an integral part of its long existence. Three years (almost) feels like a lifetime for a Tor Hidden Service.

1. Adresse accessible via TOR `http://w363zoq3ylux5rf5.onion`

## B.4 Description des dimensions de l'ACP sur Galaxy2

La figure **??** illustre l'importance des caractéristiques auprès de chacun des 5 axes de l'ACP. Les caractéristiques sont préfixées par l'aspect concerné, avec une prédominance de *med* pour média.




Poids des différentes caractéristiques sur les dimensions de l'ACP, corpus Galaxy2

# Bibliographie


[Adamcsek et al., 2006] Adamcsek, B., Palla, G., Farkas, I. J., Derényi, I., and Vicsek, T. (2006). Cfinder : locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8) :1021–1023. 57, 58, 60

[Amelio and Pizzuti, 2015] Amelio, A. and Pizzuti, C. (2015). Analysis of the italian tweet political sentiment in 2014 european elections. In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, pages 713–720. IEEE. 50, 63

[Anand et al., 2011] Anand, P., Walker, M., Abbott, R., Tree, J. E. F., Bowmani, R., and Minor, M. (2011). Cats rule and dogs drool ! : Classifying stance in online debate. In *Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis*, pages 1–9. Association for Computational Linguistics. 16, 28, 29

[Andreevskaia and Bergler, 2008] Andreevskaia, A. and Bergler, S. (2008). When specialists and generalists work together : Overcoming domain dependence in sentiment tagging. In *ACL*, pages 290–298. 29

[Asghar, 2014] Asghar, M. Z. (2014). Detection and scoring of internet slangs for sentiment analysis using sentiwordnet. *Life Science Journal*, 11(9). 23

[Aynaud and Guillaume, 2010] Aynaud, T. and Guillaume, J.-L. (2010). Static community detection algorithms for evolving networks. In *Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 proceedings of the 8th international symposium on*, pages 513–519. IEEE. 62, 63

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204. 23, 30, 70, 122

[Bahrainian and Dengel, 2013] Bahrainian, S.-A. and Dengel, A. (2013). Sentiment analysis and summarization of twitter data. In *Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on*, pages 227–234. IEEE. 26

[Bavelas, 1950] Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22(6) :725–730. 42

[Bird et al., 2009] Bird, S., Klein, E., and Loper, E. (2009). Natural language processing with python—analyzing text with the natural language toolkit o'reilly media. 68

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022. 61, 106, 123, 131, 155

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008. 55, 56, 60, 105, 111, 126, 163

[Bradley and Lang, 1999] Bradley, M. M. and Lang, P. J. (1999). Affective norms for english words (anew) : Instruction manual and affective ratings. Technical report. 14, 23, 24, 30

[Burnap et al., 2015] Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., and Sloan, L. (2015). Detecting tension in online communities with computational twitter analysis. *Technological Forecasting and Social Change*, 95 :96–108. 50

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1) :1–27. 96, 138, 161

[Caverlee et al., 2010] Caverlee, J., Liu, L., and Webb, S. (2010). The socialtrust framework for trusted social information management : Architecture and algorithms. *Information Sciences*, 180(1) :95–112. 40, 47

[Cazabet and Amblard, 2011] Cazabet, R. and Amblard, F. (2011). Simulate to detect : a multi-agent system for community detection. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, volume 2, pages 402–408. IEEE. xvii, 57

[Chen et al., 2014a] Chen, C., Gao, D., Li, W., and Hou, Y. (2014a). Inferring topic-dependent influence roles of twitter users. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1203–1206. ACM. 45, 47

[Chen et al., 2014b] Chen, M., Kuzmin, K., and Szymanski, B. K. (2014b). Community detection via maximization of modularity and its variants. *IEEE Transactions on Computational Social Systems*, 1(1) :46–65. 54

[Clauset et al., 2004] Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6) :066111. 55, 56, 60

[Cohn and Hofmann, 2001] Cohn, D. A. and Hofmann, T. (2001). The missing link-a probabilistic model of document content and hypertext connectivity. In *Advances in neural information processing systems*, pages 430–436. 61

[Costa Jr et al., 1991] Costa Jr, P. T., McCrae, R. R., and Dye, D. A. (1991). Facet scales for agreeableness and conscientiousness : A revision of the neo personality inventory. *Personality and individual Differences*, 12(9) :887–898. 32, 47

[Cunningham et al., 2011] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V., Aswani, N., Roberts, I., Gorrell, G., Funk, A., Roberts, A., Damljanovic, D., Heitz, T., Greenwood, M. A., Saggion, H., Petrak, J., Li, Y., and Peters, W. (2011). *Text Processing with GATE (Version 6)*. 29

[Damashek, 1995] Damashek, M. (1995). Gauging similarity with n-grams : Language-independent categorization of text. *Science*, 267(5199) :843. 21

[Deng and Wiebe, 2015] Deng, L. and Wiebe, J. (2015). Mpqa 3.0 : An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1323–1328. 24

[Ding et al., 2008] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM. 24

[Doan et al., 2006] Doan, A., Ramakrishnan, R., Chen, F., DeRose, P., Lee, Y., McCann, R., Sayyadian, M., and Shen, W. (2006). Community information management. *IEEE Data Eng. Bull.*, 29(1) :64–72. 50

[DuBois et al., 2009] DuBois, T., Golbeck, J., and Srinivasan, A. (2009). Rigorous probabilistic trust-inference with applications to clustering. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 655–658. IEEE. 41

[Ekman, 1992] Ekman, P. (1992). Are there basic emotions ? *Psychological Review*, 99 :550–553. 14

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6 :290–297. 51

[Farzindar and Roche, 2013] Farzindar, A. and Roche, M. (2013). Les défis de l'analyse des réseaux sociaux pour le traitement automatique des langues. *Revue TAL-Traitement Automatique des Langues*, 54(3) :7–16. 13, 18, 71

[Ferber and Gutknecht, 1998] Ferber, J. and Gutknecht, O. (1998). A meta-model for the analysis and design of organizations in multi-agent systems. In *Multi Agent Systems, 1998. Proceedings. International Conference on*, pages 128–135. IEEE. 99

[Fu et al., 2016] Fu, M., Zhu, M., Su, Y., Zhu, Q., and Li, M. (2016). Modeling temporal behavior to identify potential experts in question answering communities. In *International Conference on Cooperative Design, Visualization and Engineering*, pages 51–58. Springer. 95

[Gadek et al., 2017a] Gadek, G., Betsholtz, J., Pauchet, A., Brunessaux, S., Malandain, N., and Vercouter, L. (2017a). Extracting contextonyms from twitter for stance detection. In *ICAART (2)*, pages 132–141. 82

[Gadek et al., 2017b] Gadek, G., Pauchet, A., Malandain, N., Khelif, K., Vercouter, L., and Brunessaux, S. (2017b). Measures for topical cohesion of user communities on twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 211–218. ACM. 116

[Gadek et al., 2017c] Gadek, G., Pauchet, A., Malandain, N., Khelif, K., Vercouter, L., and Brunessaux, S. (2017c). Topical cohesion of communities on twitter. *Procedia Computer Science*, 112 :584–593. 116

[Gadek et al., 2018] Gadek, G., Pauchet, A., Malandain, N., Vercouter, L., Khelif, K., Brunessaux, S., and Grilheres, B. (2018). Topological and topical characterisation of twitter user communities. *Data Technologies and Applications*, 0(0) :20. 116

[Gaiteri et al., 2015] Gaiteri, C., Chen, M., Szymanski, B., Kuzmin, K., Xie, J., Lee, C., Blanche, T., Neto, E. C., Huang, S.-C., Grabowski, T., et al. (2015). Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering. *Scientific reports*, 5. 59, 60

[Gao et al., 2015] Gao, P., Baras, J. S., and Golbeck, J. (2015). Semiring-based trust evaluation for information fusion in social network services. In *Information Fusion (Fusion), 2015 18th International Conference on*, pages 590–596. IEEE. 41

[Gatti et al., 2013] Gatti, M., Cavalin, P., Neto, S. B., Pinhanez, C., dos Santos, C., Gribel, D., and Appel, A. P. (2013). Large-scale multi-agent-based modeling and simulation of microblogging-based online social network. In *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, pages 17–33. Springer. 36, 47

[Giatsidis et al., 2013] Giatsidis, C., Malliaros, F. D., and Vazirgiannis, M. (2013). Advanced graph mining for community evaluation in social networks and the web. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 771–772. ACM. 58, 60

[Gilpin et al., 2013] Gilpin, S., Eliassi-Rad, T., and Davidson, I. (2013). Guided learning for role discovery (glrd) : framework, algorithms, and applications. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 113–121. ACM. 44, 47

[Giorgio Bertolin, 2017] Giorgio Bertolin, N. S. C. C. o. E. (2017). *Digital Hydra : Security Implications of False Information Online*. NATO StratCom COE. 5

[Girvan and Newman, 2002] Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12) :7821–7826. 56

[Golbeck and Hendler, 2004] Golbeck, J. and Hendler, J. (2004). Accuracy of metrics for inferring trust and reputation in semantic web-based social networks. *Engineering knowledge in the age of the semantic web*, pages 116–131. 41

[Goldenberg et al., 2001] Goldenberg, J., Libai, B., and Muller, E. (2001). Talk of the network : A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12(3) :211–223. 38, 47

[Gotti et al., 2013] Gotti, F., Langlais, P., and Farzindar, A. (2013). Translating government agencies' tweet feeds : Specificities, problems and (a few) solutions. *NAACL 2013*, page 80. 68

[Granovetter, 1978] Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6) :1420–1443. 38, 47

[Greene et al., 2012] Greene, D., O'Callaghan, D., and Cunningham, P. (2012). Identifying topical twitter communities via user list aggregation. *COMMPER 2012*, page 41. 60

[Gregory, 2007] Gregory, S. (2007). An algorithm to find overlapping community structure in networks. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 91–102. Springer. 58

[Gregory, 2008] Gregory, S. (2008). A fast algorithm to find overlapping communities in networks. *Machine learning and knowledge discovery in databases*, pages 408–423. 58, 60

[Gregory, 2010] Gregory, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10) :103018. 58, 60

[Guille and Favre, 2015] Guille, A. and Favre, C. (2015). Event detection, tracking, and visualization in twitter : a mention-anomaly-based approach. *Social Network Analysis and Mining*, 5(1) :1–18. 47, 61

[Guille et al., 2013] Guille, A., Hacid, H., Favre, C., and Zighed, D. A. (2013). Information diffusion in online social networks : A survey. *ACM SIGMOD Record*, 42(2) :17–28. 38

[Hagberg et al., 2008] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA USA. 75

[Hasan and Ng, 2013] Hasan, K. S. and Ng, V. (2013). Extra-linguistic constraints on stance recognition in ideological debates. In *ACL (2)*, pages 816–821. 29, 30, 70

[Henderson et al., 2012] Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Faloutsos, C., and Li, L. (2012). Rolx : structural role extraction & mining in large graphs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1231–1239. ACM. 44, 47, 98, 99, 124, 140

[Henderson et al., 2011] Henderson, K., Gallagher, B., Li, L., Akoglu, L., Eliassi-Rad, T., Tong, H., and Faloutsos, C. (2011). It's who you know : graph mining using recursive structural features. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 663–671. ACM. 44

[Herzig et al., 2009] Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L. (2009). A logic of trust and reputation. *Logic Journal of IGPL*, 18(1) :214–244. 40

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1) :193–218. 54

[Hutto and Gilbert, 2014] Hutto, C. J. and Gilbert, E. (2014). Vader : A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*. 24, 30, 122, 156

[Hyungsuk et al., 2003] Hyungsuk, J., Ploux, S., and Wehrli, E. (2003). Lexical knowledge representation with contexonyms. In *9th MT summit Machine Translation*, pages 194–201. 72

[Jaccard, 1901] Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, 37 :547–579. 54

[Jendoubi et al., 2017] Jendoubi, S., Martin, A., Liétard, L., Hadji, H. B., and Yaghlane, B. B. (2017). Two evidential data based models for influence maximization in twitter. *Knowledge-Based Systems*, 121 :58–70. 39

[Jiang et al., 2014] Jiang, W., Wang, G., and Wu, J. (2014). Generating trusted graphs for trust evaluation in online social networks. *Future generation computer systems*, 31 :48–58. 42

[Jøsang et al., 2007] Jøsang, A., Ismail, R., and Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2) :618–644. 40

[Kalchbrenner et al., 2014] Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, volume 1, pages 655–665. 22

[Kempe et al., 2003] Kempe, D., Kleinberg, J., and Tardos, É. (2003). Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM. 39

[Khan et al., 2015] Khan, A. Z., Atique, M., and Thakare, V. (2015). Combining lexicon-based and learning-based methods for twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, page 89. xvii, 26, 27

[Khan et al., 2014] Khan, F. H., Bashir, S., and Qamar, U. (2014). Tom : Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, 57 :245–257. xvii, 24, 25, 30

[Kuhn, 1955] Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2) :83–97. 54

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media ? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM. 42, 45, 47, 50, 60, 89, 138, 160, 161

[Lagnier et al., 2013] Lagnier, C., Denoyer, L., Gaussier, E., and Gallinari, P. (2013). Predicting information diffusion in social networks using content and user's profiles. In *European conference on information retrieval*, pages 74–85. Springer. 39

[Lancichinetti et al., 2011] Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, 6(4) :e18961. 59, 60

[Lasswell, 1948] Lasswell, H. D. (1948). The structure and function of communication in society. *New York*. 5, 31

[Le and Mikolov, 2014] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196. 22

[Lee et al., 2010] Lee, C., Kwak, H., Park, H., and Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th international conference on World wide web*, pages 1137–1138. ACM. 45, 47, 160

[Lim and Datta, 2012] Lim, K. H. and Datta, A. (2012). Finding twitter communities with common interests using following links of celebrities. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 25–32. ACM. 50, 63

[Lim and Datta, 2016] Lim, K. H. and Datta, A. (2016). An interaction-based approach to detecting highly interactive twitter communities using tweeting links. In *Web Intelligence*, volume 14, pages 1–15. IOS Press. 60, 104, 115

[Liu, 2010] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2 :627–666. 15, 25

[Liu et al., 2010] Liu, J., Dolan, P., and Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM. 35

[Liu et al., 2015] Liu, P., Joty, S., and Meng, H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*. xix, 27, 28, 30

[Liu et al., 2014] Liu, S., Wang, S., Zhu, F., Zhang, J., and Krishnan, R. (2014). Hydra : Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 51–62. ACM. xvii, 34, 47

[Liu et al., 2009] Liu, Y., Niculescu-Mizil, A., and Gryc, W. (2009). Topic-link lda : joint models of topic and author community. In *proceedings of the 26th annual international conference on machine learning*, pages 665–672. ACM. 61

[Maynard et al., 2012] Maynard, D., Bontcheva, K., and Rout, D. (2012). Challenges in developing opinion mining tools for social media. *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pages 15–22. 18, 68

[Maynard and Funk, 2011] Maynard, D. and Funk, A. (2011). Automatic detection of political opinions in tweets. In *The semantic web : ESWC 2011 workshops*, pages 88–99. Springer. 29, 30

[Mendoza et al., 2010] Mendoza, M., Poblete, B., and Castillo, C. (2010). Twitter under crisis : Can we trust what we rt ? In *Proceedings of the first workshop on social media analytics*, pages 71–79. ACM. 40

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119. 22, 75

[Miller, 1995] Miller, G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, 38(11) :39–41. 23, 75

[Mohamadi-Baghmolaei et al., 2015] Mohamadi-Baghmolaei, R., Mozafari, N., and Hamzeh, A. (2015). Trust based latency aware influence maximization in social networks. *Engineering Applications of Artificial Intelligence*, 41(C) :195–206. 40

[Mohammad et al., 2016] Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016). Semeval-2016 task 6 : Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval*, volume 16. 16, 80, 81

[Mohammad and Turney, 2010] Mohammad, S. M. and Turney, P. D. (2010). Emotions evoked by common words and phrases : Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics. 15, 23, 24, 30

[Munezero et al., 2015] Munezero, M., Montero, C. S., Mozgovoy, M., and Sutinen, E. (2015). Emotwitter–a fine-grained visualization system for identifying enduring sentiments in tweets. In *Computational Linguistics and Intelligent Text Processing*, pages 78–91. Springer. 15, 30

[Nakov et al., 2013] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2 : Sentiment analysis in twitter. 26

[Nallapati et al., 2008] Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550. ACM. 61

[NATO, 2016] NATO, S. C. C. o. E. (2016). *Daesh information campaign and its influence*. NATO StratCom COE. 4

[Nepal et al., 2011] Nepal, S., Sherchan, W., and Paris, C. (2011). Strust : A trust model for social networks. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on*, pages 841–846. IEEE. 41

[Newman, 2006] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23) :8577–8582. 51

[Newman and Girvan, 2004] Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2) :026113. 43, 53, 55, 60

[Newman et al., 2001] Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2) :026118. 51

[Nicosia et al., 2008] Nicosia, V., Mangioni, G., Malgeri, M., and Carchiolo, V. (2008). Extending modularity definition for directed graphs with overlapping communities. Technical report. 53

[Noordhuis et al., 2010] Noordhuis, P., Heijkoop, M., and Lazovik, A. (2010). Mining twitter in the cloud : A case study. In *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, pages 107–114. IEEE. 45, 47

[Nystuen and Dacey, 1961] Nystuen, J. D. and Dacey, M. F. (1961). A graph theory interpretation of nodal regions. In *Papers of the Regional Science Association*, volume 7, pages 29–42. Springer. 55, 56, 60

[Onnela et al., 2007] Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the national academy of sciences*, 104(18) :7332–7336. 51

[Owoputi et al., 2013] Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. Association for Computational Linguistics. 21, 73

[Page et al., 1999] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking : bringing order to the web. 43, 89

[Pak and Paroubek, 2010] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326. 25, 30

[Palla et al., 2005] Palla, G., Derényi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435 :814–818. 50, 58, 74, 75

[Panek et al., 2013] Panek, E. T., Nardis, Y., and Konrath, S. (2013). Mirror or megaphone ? : How relationships between narcissism and social networking site use differ on facebook and twitter. *Computers in Human Behavior*, 29(5) :2004–2012. 32

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2) :1–135. 15, 25, 28, 30

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics. 15, 30

[Pearce et al., 2014] Pearce, W., Holmberg, K., Hellsten, I., and Nerlich, B. (2014). Climate change on twitter : Topics, communities and conversations about the 2013 ipcc working group 1 report. *PloS one*, 9(4) :e94785. 62

[Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, 12 :2825–2830. 70

[Pennebaker et al., 2001] Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count : Liwc 2001. *Mahway : Lawrence Erlbaum Associates*, 71 :2001. 23, 24, 123

[Picard, 1995] Picard, W. R. (1995). Affective computing. *Technical Report 321 MIT Media Laboratory*, page 24. 14

[Ploux and Hyungsuk, 2003] Ploux, S. and Hyungsuk, J. (2003). A model for matching semantic maps between languages (french/english, english/french). *Computational linguistics*, 29(2) :155–178. 72

[Plutchik, 1980] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of emotion*, 1 :3–31. 14, 24

[Pons and Latapy, 2005] Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences*, pages 284–293. Springer. 55, 57, 60

[Poria et al., 2016] Poria, S., Cambria, E., and Gelbukh, A. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, 108 :42–49. 28, 30

[Quercia et al., 2011] Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves : Predicting personality with twitter. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 180–185. IEEE. 32, 47

[Queyroi et al., 2015] Queyroi, F., Beauguitte, L., and Pecout, H. (2015). Rss flows, world structure & community detection. In *European Colloquium of Theoretical and Quantitative Geography*. 56, 58

[Raghavan et al., 2007] Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3) :036106. 55, 57, 58, 60

[Raghavan et al., 2014] Raghavan, V., Ver Steeg, G., Galstyan, A., and Tartakovsky, A. G. (2014). Modeling temporal activity patterns in dynamic social networks. *IEEE Transactions on Computational Social Systems*, 1(1) :89–107. xvii, 36, 37, 47

[Rao et al., 2015] Rao, A., Spasojevic, N., Li, Z., and DSouza, T. (2015). Klout score : Measuring influence across multiple social networks. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2282–2289. IEEE. 46, 47, 88

[Reed and Kadayam, 2017] Reed, M. and Kadayam, S. (2017). Topical trust network. US Patent 9,607,324. 41

[Rosa et al., 2011] Rosa, K. D., Shah, R., Lin, B., Gershman, A., and Frederking, R. (2011). Topical clustering of tweets. *Proceedings of the ACM SIGIR : SWSM*. 63

[Rosvall and Bergstrom, 2007] Rosvall, M. and Bergstrom, C. (2007). Maps of information flow reveal community structure in complex networks. Technical report, Technical report. 55, 56, 60

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer. 21

[Şerban, 2013] Şerban, O. (2013). *Detection and integration of affective feedback into distributed interactive systems*. PhD thesis, Citeseer. 72

[Şerban et al., 2012] Şerban, O., Pauchet, A., Rogozan, A., Pécuchet, J.-P., and LITIS, I. (2012). Semantic propagation on contextonyms using sentiwordnet. In *WACAI 2012 Workshop Affect, Compagnon Artificiel, Interaction*, page 86. 23, 73

[Shaheen, 2015] Shaheen, J. (2015). *Network of terror : how DAESH uses adaptive social networks to spread its message*. Nato StratCom CoE. 5

[Sherchan et al., 2013] Sherchan, W., Nepal, S., and Paris, C. (2013). A survey of trust in social networks. *ACM Computing Surveys (CSUR)*, 45(4) :47. 40, 47

[Shi and Malik, 2000] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8) :888–905. 53

[Sievert and Shirley, 2014] Sievert, C. and Shirley, K. E. (2014). Ldavis : A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70. 156

[Socher et al., 2013] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642. xvii, 26

[Strapparava et al., 2004] Strapparava, C., Valitutti, A., et al. (2004). Wordnet affect : an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086. 23, 30

[Subrahmanian et al., 2016] Subrahmanian, V., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F., et al. (2016). The darpa twitter bot challenge. *arXiv preprint arXiv :1601.05140*. 33, 47

[Tan et al., 2002] Tan, C.-M., Wang, Y.-F., and Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Information processing & management*, 38(4) :529–546. 21

[Tao et al., 2011] Tao, K., Abel, F., Gao, Q., and Houben, G.-J. (2011). Tums : twitter-based user modeling service. In *Extended Semantic Web Conference*, pages 269–283. Springer. 32

[Tsytsarau and Palpanas, 2012] Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3) :478–514. 28, 30

[Tyshchuk et al., 2014] Tyshchuk, Y., Wallace, W. A., Li, H., Ji, H., and Kase, S. E. (2014). The nature of communications and emerging communities on twitter following the 2013 syria sarin gas attacks. In *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, pages 41–47. IEEE. 60

[Van Dongen, 2000] Van Dongen, S. M. (2000). *Graph clustering by flow simulation*. PhD thesis. 55, 57, 60

[Varol et al., 2014] Varol, O., Ferrara, E., Ogan, C. L., Menczer, F., and Flammini, A. (2014). Evolution of online user behavior during a social upheaval. In *Proceedings of the 2014 ACM conference on Web science*, pages 81–90. ACM. 94

[Wang et al., 2016] Wang, R., Zhao, H., Ploux, S., Lu, B.-L., and Utiyama, M. (2016). A bilingual graph-based semantic model for statistical machine translation. In *International Joint Conference on Artificial Intelligence*. 72

[Wang et al., 2015] Wang, Y., Li, L., and Liu, G. (2015). Social context-aware trust inference for trust enhancement in social network based recommendations on service providers. *World Wide Web*, 18(1) :159–184. 41

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, 393(6684) :440. 51

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2) :165–210. 24, 30

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics. 23, 24

[Xie et al., 2013] Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks : The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4) :43. 59

[Xie and Szymanski, 2012] Xie, J. and Szymanski, B. K. (2012). Towards linear time overlapping community detection in social networks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 25–36. Springer. 59, 60

[Yamak et al., 2018] Yamak, Z., Saunier, J., and Vercouter, L. (2018). Sockscatch : Automatic detection and grouping of sockpuppets in social media. *Knowledge-Based Systems*, 149 :124–142. 35

[Yang and Leskovec, 2015] Yang, J. and Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1) :181–213. 52

[Yang et al., 2014] Yang, Z., Wilson, C., Wang, X., Gao, T., Zhao, B. Y., and Dai, Y. (2014). Uncovering social network sybils in the wild. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(1) :2. 33

[Yassine and Hajj, 2010] Yassine, M. and Hajj, H. (2010). A framework for emotion mining from text in online social networks. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1136–1142. IEEE. 15, 30

[Yin et al., 2014] Yin, H., Cui, B., Chen, L., Hu, Z., and Huang, Z. (2014). A temporal context-aware model for user behavior modeling in social media systems. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1543–1554. ACM. 33

[Yin et al., 2016] Yin, H., Cui, B., Zhou, X., Wang, W., Huang, Z., and Sadiq, S. (2016). Joint modeling of user check-in behaviors for real-time point-of-interest recommendation. *ACM Transactions on Information Systems (TOIS)*, 35(2) :11. 35

[Yin et al., 2012] Yin, Z., Cao, L., Gu, Q., and Han, J. (2012). Latent community topic analysis : Integration of community discovery with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4) :63. 54, 61

[Yuan et al., 2013] Yuan, Q., Cong, G., Ma, Z., Sun, A., and Thalmann, N. M. (2013). Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 363–372. ACM. 36

[Zamparas et al., 2015] Zamparas, V., Kanavos, A., and Makris, C. (2015). Real time analytics for measuring user influence on twitter. In *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, pages 591–597. IEEE. 45

[Zhou et al., 2006] Zhou, D., Manavoglu, E., Li, J., Giles, C. L., and Zha, H. (2006). Probabilistic models for discovering e-communities. In *Proceedings of the 15th international conference on World Wide Web*, pages 173–182. ACM. 61

## Détection d'opinions, d'acteurs-clés et de communautés thématiques dans les médias sociaux

Les réseaux sociaux numériques ont pris une place prépondérante dans l'espace informationnel. Alors que la quantité d'information rend difficile son exploitation par des humains, le besoin reste entier d'analyser un réseau social numérique : il faut dégager des tendances à partir des messages postés, qualifier les comportements des utilisateurs, et identifier les structures sociales émergentes.

Nous proposons un système d'analyse en trois niveaux. Tout d'abord l'analyse du message vise à en déterminer l'opinion. Ensuite, la caractérisation et l'évaluation des comptes utilisateurs est réalisée grâce à une étape de profilage comportemental et à l'étude de leur position dans des graphes sociaux, combinées avec les statistiques d'engagement, par exemple en nombre d'abonnés. Enfin, le système détecte et évalue des communautés d'utilisateurs, pour lesquelles nous introduisons des scores de cohésion thématique qui complètent les mesures topologiques classiques de qualité structurelle des communautés détectées.

## Detection of opinions, key-actors and thematic communities in online social media

Online Social Networks have taken a huge place in the informational space, and are often used for advertising, e-reputation, propaganda, or even manipulation, either by individuals, companies or states. As the quantity of information makes difficult the human exploitation, the need for social network analysis remains unsatisfied: trends must be extracted from the posted messages, the user behaviours must be characterised, and the social structure must be identified.

To tackle this problem, we propose a system providing analysis tools on three levels, beginning with message analysis, to determine the opinions they bear. Next, user account characterisation and evaluation is performed thanks to a behavioural profiling method, the study of node importance and position in social graphs, and engagement and influence measures. Finally the system proceeds to user community detection and evaluation. For this last challenge, we introduce thematic cohesion scores, completing the topological, graph-based measures for group quality.