Modélisation individuelle et multi-factorielle du phénomène de polarisation pour une personnalisation de l'apport en diversité dans les recommandations de news

Céline Treuillier

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Modélisation individuelle et multi-factorielle du phénomène de polarisation pour une personnalisation de l'apport en diversité dans les recommandations de news

THÈSE

présentée et soutenue publiquement le 18 octobre 2024

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Céline Treuillier

Composition du jury

Rapporteurs :	**Camille Salinesi** Professeur des Universités, Université Paris 1 Panthéon - Sorbonne
	Elsa Negre Maîtresse de conférences HDR, Université Paris - Dauphine
Présidente :	**Karën Fort** Professeure des Universités, Université de Lorraine
Examinateur :	**Raphaël Fournier-S'niehotta** Maître de conférences HDR, CNAM Paris
Encadrants :	**Armelle Brun** Professeure des Universités, Université de Lorraine
	Sylvain Castagnos Maître de conférences, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Mis en page avec la classe thesul.

Remerciements

Je voudrais tout d'abord remercier l'ensemble des membres du jury pour leur intérêt pour ce travail de thèse. En particulier, je remercie Camille Salinesi et Elsa Negre pour avoir accepté d'être rapporteurs, ainsi que Karën Fort et Raphaël Fournier-S'niehotta pour avoir accepté d'être examinateurs. Les rapports, retours et discussions durant la soutenance ont été très enrichissants et ont notamment permis de mettre en lumière divers pistes de recherches intéressantes.

Par ailleurs, bien que ce manuscrit porte mon nom, le travail qui y est présenté est le fruit d'un travail collaboratif. Durant mes trois années de thèse, j'ai eu l'opportunité d'échanger et de travailler avec des chercheurs de tous horizons, dont les contributions ont été précieuses et ont grandement participé à la réussite de mes travaux.

Je souhaite tout d'abord exprimer mes sincères remerciements à ma directrice de thèse, Armelle Brun. Sa bienveillance, la qualité de ses conseils et sa disponibilité, depuis mon entrée en Master en Sciences Cognitives jusqu'à la fin de ma thèse, sont inestimables. Armelle m'a transmis sa passion pour la recherche et a eu un impact considérable sur mon parcours universitaire. Je remercie également chaleureusement mon co-directeur de thèse, Sylvain Castagnos, pour son soutien et ses précieuses suggestions, qui ont grandement contribué à améliorer la qualité de ce travail. Je suis très reconnaissante d'avoir bénéficié d'un encadrement d'une telle qualité, tant sur le plan scientifique qu'humain.

Je tiens par ailleurs à remercier tous ceux qui ont contribué, de près ou de loin à la réussite de mon travail. Je pense évidemment à Anne Boyer, qui m'a offert l'opportunité de rejoindre le monde de la recherche lors de ma seconde année de Master. Travailler à ses côtés durant une année a considérablement renforcé mon envie de poursuivre mon parcours universitaire par une thèse. Je remercie également Özlem Özgöbek pour m'avoir accueillie dans son équipe à l'Université Norvégienne de Sciences et de Technologie. Ce séjour de recherche à l'étranger a été très enrichissant et m'a permis de travailler dans un nouvel environnement de recherche très riche.

Je saisis cette occasion pour exprimer mes remerciements aux membres du projet BOOM, avec qui j'ai eu le plaisir de collaborer et de partager des moments agréables entre Avignon, Paris et Nancy. Je remercie aussi tous les membres de l'équipe BIRD pour leurs conseils avisés et leur bonne humeur. Je suis profondément reconnaissante d'avoir évolué dans un environnement scientifique de haute qualité, entourée de personnes qui m'ont permis de progresser et de développer un projet professionnel qui me passionne.

Finalement, je remercie évidemment ma famille, et particulièrement mes parents et mon frère. J'espère avoir su exprimer toute ma gratitude pour votre soutien indéfectible tout au long de mon parcours. Merci à mes amis, avec qui je partage des moments inoubliables depuis près de 20 ans pour certains, et seulement quelques mois pour d'autres. Vous occupez tous une place très importante dans ma vie, et je suis profondément reconnaissante d'être si bien entourée. Je tiens à remercier tout particulièrement Étienne pour son soutien quotidien.

Table des matières

Table des figures

Liste des tableaux

Introduction

Depuis les années 2000, la démocratisation de l'utilisation d'Internet et des réseaux sociaux s'est accompagnée d'une profonde modification du paysage informationnel et des pratiques associées. L'ubiquité des outils numériques a durablement transformé la manière dont les informations sont diffusées, mais aussi la façon dont elles sont consommées. Les médias traditionnels comme la télévision ou la radio cohabitent désormais avec les plateformes d'information en ligne et les réseaux sociaux, accessibles depuis téléphones mobiles, tablettes et ordinateurs. En 2023, le taux de pénétration d'Internet est de 92 % en France, et 69 % des français visitent chaque mois les sites d'actualité en ligne pour s'informer [1]. Cette accessibilité permanente, immédiate et multimodale à l'information offre à chacun les mêmes opportunités de s'informer et de s'engager activement dans le débat public. Dans un système politique démocratique, cet accès universalisé à l'information revêt une importance cruciale. Les médias y jouent donc un rôle important, en alimentant les discussions à travers la publication quotidienne de nouvelles, ou *news*, couvrant des thématiques diverses et offrant de multiples perspectives. Cette pluralité de sources d'information, de points de vue et de sujets contribue à créer un paysage médiatique dynamique et diversifié.

Ce paysage informationnel transformé s'accompagne évidemment d'une production massive de données : des millions de contenus sont publiés et rendus disponibles chaque jour dans le monde [2]. Des technologies doivent alors être développées pour assurer le traitement et le stockage de cette vaste masse de données, appelée *big data*. Cette prolifération de contenus peut également entraîner une surcharge d'information, rendant difficile pour les individus d'identifier les informations qui leurs sont pertinentes et de discerner les faits des opinions [Schmitt *et al.*, 2018]. En effet, les capacités cognitives humaines ne permettent pas de traiter cette surabondance d'information [Golman *et al.*, 2017]. L'accès aux différents contenus disponibles afin de trouver celui d'intérêt est évidemment inenvisageable tant en termes de temps requis que de capacités cognitives. Ces limites reflètent le concept de rationalité limitée décrit par Herbert Simon [Simon, 1957], selon lequel les décisions des individus sont entravées par un ensemble de contraintes, les incitant ainsi à privilégier des solutions satisfaisantes plutôt qu'optimales. Pour aider les utilisateurs dans leur quête informationnelle, des systèmes doivent ainsi être mis en place pour trouver le contenu d'intérêt parmi la vaste quantité d'informations disponibles. Un **filtrage personnalisé et automatisé de l'information** devient alors essentiel [Baeza-Yates *et al.*, 1999].

L'Intelligence Artificielle (IA) permet de répondre à ce besoin de filtrage de l'information en permettant une prise de décision automatisée par l'application de méthodes de résolutions de problèmes. Dans le domaine de la recherche d'information, les moteurs de recherche traditionnels ont par exemple longtemps dominé, répondant aux requêtes des utilisateurs par des résultats filtrés sur la base de mots-clés ou de commandes vocales [Manning *et al.*, 2008]. Ces systèmes, bien que

1. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2023
2. https://www.pewresearch.org/journalism/fact-sheet/newspapers/

fondamentaux, se heurtent à une limite importante : la nécessité et la capacité pour l'utilisateur d'exprimer explicitement un besoin. Pour répondre à cette contrainte, les **systèmes de recommandation** ont émergé, exploitant les besoins implicites des utilisateurs plutôt que des requêtes formulées de manière explicite [Resnick and Varian, 1997]. Ces systèmes innovants s'appuient sur **l'exploitation et l'analyse des interactions des utilisateurs en ligne pour modéliser et anticiper leurs besoins**. Pour cela, les algorithmes d'IA appliquent des approches d'apprentissage automatique permettant de filtrer les informations. Certains adoptent une analyse collaborative, appelée *filtrage collaboratif* [Resnick *et al.*, 1994], comparant les utilisateurs entre eux pour trouver des intérêts communs. D'autres, appliquent une *recommandation basée contenu* [Pazzani and Billsus, 2007] en exploitant les métadonnées et les contenus pour établir des correspondances entre les intérêts des utilisateurs et les items disponibles. Cette distinction entre systèmes de filtrage collaboratif et systèmes basés contenu n'est pas exclusive, et certains tirent profit des avantages de chacune des approches dans des *systèmes hybrides* [Burke, 2002].

Avec l'avènement d'Internet et l'essor des technologies numériques, les systèmes de recommandation sont devenus omniprésents, influençant notre consommation de contenu numérique au quotidien [Jannach and Zanker, 2012]. Ils se manifestent sous diverses formes, des recommandations de films et de séries sur les services de streaming, aux suggestions de traitement médicamenteux, en passant par le filtrage du contenu informatif sur les plateformes d'actualité en ligne ou les réseaux sociaux. Dans le domaine spécifique de **la recommandation de news**, l'objectif est d'offrir un filtrage automatisé et personnalisé de l'actualité. Cela implique l'identification des news susceptibles d'intéresser l'utilisateur en se basant sur une analyse approfondie de ses préférences implicites et de ses habitudes passées de consultation de l'information.

Quel que soit leur mode de fonctionnement et le contexte d'application, les systèmes de recommandation sont constamment affinés pour améliorer leur précision, correspondant à leur capacité à correctement identifier les items pertinents pour les utilisateurs [Konstan *et al.*, 1998, Bellogin *et al.*, 2011]. Cette recherche de performance centrée sur la précision des résultats a cependant montré ses limites. En cherchant à maximiser cette précision, un phénomène de surspécialisation (ou *"over-specialization"*) s'installe [McNee *et al.*, 2006]. Dans le domaine de la recommandation de news, l'optimisation exclusive de la précision entraîne la recommandation de news similaires à celles préalablement consultées, et de surcroît un manque important de diversité dans le contenu recommandé. En 2011, Eli Pariser introduit la notion de *"bulles de filtre"*, au sein desquelles les utilisateurs se retrouvent isolés sous l'action d'algorithmes de filtrage, et où ils ne sont exposés qu'à une sélection d'informations spécifiques, non représentative de la diversité des points de vue disponibles. Dans cet environnement clos, l'accès à une variété d'opinions et d'idées est artificiellement limité, sans que les utilisateurs en aient conscience [Pariser, 2011]. Cet enfermement soulève également des enjeux importants liés à la manière dont les individus reçoivent et traitent l'information dans une ère numérique dominée par l'IA et les algorithmes de filtrage [Spohr, 2017]. La personnalisation des recommandations est notamment susceptible d'involontairement limiter l'exposition à des idées diverses, renforçant ainsi les préjugés existants et contribuant au renforcement du phénomène de **polarisation des opinions** [Kubin and Von Sikorski, 2021]. Reflétant une prise de position forte et le rejet des opinions adverses, cette polarisation peut favoriser l'adoption de comportements extrêmes et fragiliser l'équilibre démocratique des sociétés contemporaines.

Cependant, bien que revêtant cette caractéristique de dangerosité, cette polarisation fait débat dans les discours scientifiques [Sunstein, 1999]. Nulle définition commune n'a pu être établie, et tandis que certains prônent le besoin de l'éviter, d'autres mettent en avant le besoin de la conserver, de façon raisonnée, pour le maintien d'un débat démocratique sain [McCoy *et al.*, 2018].

Elle attire des chercheurs de disciplines multiples : sciences sociales, sciences politiques, psychologie, mathématiques, informatique, *etc.* Certains tentent d'en décrire les processus sous-jacents [Levin *et al.*, 2021], d'en modéliser le fonctionnement [Sîrbu *et al.*, 2017], d'étudier les répercussions exactes sur la société [Milačić, 2021] ou encore de participer à son contrôle ou à sa réduction [Lorenz-Spreen *et al.*, 2020]. **Dans le domaine de l'IA, cela inclut le développement de systèmes de recommandation plus transparents et éthiques [Trattner *et al.*, 2022], encourageant l'exposition à un spectre plus large d'informations tout en respectant les préférences et la liberté de choix des utilisateurs.** Ces systèmes héritent alors d'un rôle démocratique [Helberger, 2021].

En pratique, cela nécessite d'intégrer des approches automatiques favorisant la découverte et l'exploration dans les algorithmes de recommandation [Helberger *et al.*, 2018]. Il s'agit notamment d'une introduction délibérée de diversité dans les recommandations ou d'une promotion de contenus qui défient les préférences établies de l'utilisateur. En fin de compte, l'objectif est de développer des systèmes répondant aux intérêts immédiats de l'utilisateur (*précision*), tout en encouragent l'apprentissage et la découverte continue (*diversité*). Cela reflète un défi fondamental de tout système de recommandation : **apporter de la diversité dans les recommandations tout en assurant un équilibre avec la personnalisation de ces recommandations**. Dans le domaine spécifique de la recommandation de news, les chercheurs doivent répondre à une double exigence : concevoir des approches de sélection de l'information capables d'identifier les news pertinentes, mais susceptibles de renforcer la polarisation des opinions, et concevoir des systèmes capables de mesurer et de valoriser la diversité des contenus proposés pour contrer le phénomène de polarisation, au risque de ne pas satisfaire les utilisateurs.

1 Motivations et questions de recherche

Avec l'évolution des habitudes de consommation de l'information, il est crucial d'adapter les systèmes de filtrage automatisé et personnalisé de l'information. **L'impact social de la recommandation de news souligne l'importance de l'apport en diversité, et de la considération d'aspects sociaux et sociétaux par les algorithmes de filtrage.** Cependant, le développement d'algorithmes répondant aux besoins des utilisateurs tout en promouvant la diversité est un défi de taille. Bien que l'introduction de diversité dans les recommandations de news semble être une solution prometteuse pour combattre la polarisation, son acceptation par les utilisateurs et la manière dont elle influence réellement leurs habitudes de consommation d'informations reste un sujet d'étude ouvert.
La problématique ayant guidé mon travail de thèse est donc la suivante :

Problématique
Comment participer à réduire le phénomène de polarisation en ligne en confrontant les utilisateurs à un contenu plus diversifié ?

L'impact de cet apport en diversité dans les recommandations de news sur l'acceptation et la consommation effective des utilisateurs est source de questionnement. Cela m'a notamment amenée à poser une hypothèse selon laquelle la diversification des recommandations n'est pas accueillie de façon universelle par l'ensemble des utilisateurs. Cette hypothèse a pu être validée au travers d'une étude préliminaire, non détaillée dans ce manuscrit, ayant confirmé qu'un apport en diversité n'a pas nécessairement un impact positif sur la consommation de l'actualité. Ces résultats viennent compléter les conclusions selon lesquelles une confrontation à un

contenu divers peut non seulement faillir à réduire la polarisation, mais parfois même la renforcer [Bail *et al.*, 2018]. Par ailleurs, la distinction de différents profils d'utilisateurs permet de confirmer le besoin d'approches de recommandation et de diversification qui soient personnalisées et adaptées aux besoins de chaque utilisateur. Finalement, cette étude a également permis de confirmer que l'impact de l'apport en diversité est susceptible de fluctuer au cours du temps [Treuillier *et al.*, 2022].

Ces conclusions révèlent un manque de compréhension approfondie de la polarisation en ligne, ainsi que des comportements et des dynamiques qui lui sont associés. De mon point de vue, ce manque découle notamment d'une modélisation trop superficielle de la polarisation présentant trois limites majeures : (1) elle repose sur une exploitation partielle des données, ne tenant donc pas compte des différents facteurs associés ; (2) elle adopte une approche globale de la polarisation, et ne modélise pas les comportements individuels ; (3) elle ne considère jamais les variations temporelles. Par conséquent, les approches de diversification mises en œuvre sont inadaptées, et leur véritable impact sur la consommation d'actualités demeure incertain. Nous soutenons alors ici l'hypothèse forte selon laquelle **une modélisation à la fois individuelle, multi-factorielle et temporelle du phénomène de polarisation en ligne est essentielle au développement d'approches de diversification personnalisées et adaptées dans le domaine de la recommandation de news**. Une première question de recherche se pose alors :

> ***Question de Recherche 1 (QR1)***
> Comment modéliser les comportements de polarisation individuels de façon multi-factorielle pour tenir compte de la complexité du phénomène de polarisation, et de façon temporelle pour rendre compte de la dynamique sous-jacente ?

Bien que l'apport en diversité adapté et personnalisé soit l'un des objectifs principaux de cette thèse, le respect des préférences des utilisateurs reste primordial pour assurer l'acceptation du système. S'y ajoute alors l'ensemble des dimensions éthiques liées à la recommandation de news, cette dernière ayant un fort impact sur la société. Au cœur d'une industrie médiatique complexe, les concepteurs de systèmes d'IA doivent veiller à ce que les systèmes, tels qu'implémentés et entraînés, ne perpétuent pas les biais existants ou n'introduisent pas de nouvelles formes de discrimination. **L'apport en diversité ne doit en effet pas être appliqué au détriment de la précision et de l'éthique, et ne doit pas non plus orienter artificiellement les opinions des utilisateurs.** Cela implique une réflexion approfondie sur les valeurs que ces systèmes sont censés promouvoir et la mise en place de garde-fous pour prévenir les biais et les impacts non contrôlés. Ceci amène à la question de recherche suivante :

> ***Question de Recherche 2 (QR2)***
> Comment modifier les approches de recommandation de news de façon à apporter de la diversité tout en respectant les préférences des utilisateurs et les aspects éthiques ?

L'entreprise d'une telle tâche n'est pas nouvelle. De nombreux chercheurs et experts des systèmes de recommandation s'y sont attelés. Cependant, cette quête de réponse aux défis de notre société moderne est parfois obscurcie par une course à la performance. En effet, la discipline vit une époque marquée par l'avènement des systèmes profonds, qui permettent d'atteindre des performances jamais obtenues jusqu'alors. Ces systèmes toujours plus complexes et puissants occupent désormais le devant de la scène, et sont largement plébiscités pour leurs exploits. La démocratisation d'outils comme ChatGPT en est un exemple concret, montrant à quel point ces systèmes sont capables de modifier les habitudes [Kalla and Smith, 2023]. Dans le contexte

du développement de ces systèmes puissants, le besoin d'une compréhension approfondie des phénomènes comportementaux et sociaux sous-jacents à la polarisation se fait parfois timide. Les travaux résultants de collaborations inter-disciplinaires, nécessaires lors du travail sur de telles thématiques, perdent ainsi parfois en visibilité et en reconnaissance. Bien loin de nier les performances des systèmes profonds, le travail réalisé durant cette thèse repose néanmoins sur l'idée qu'une recherche informée par l'expertise des disciplines connexes comme la psychologie ou les sciences politiques permet de proposer des solutions plus frugales tout en restant pertinentes et efficaces. La thèse présentée dans ce manuscrit n'en reste évidemment pas moins une thèse en informatique, abordant la modélisation utilisateur et les systèmes de recommandation.

2 Contributions

Pour répondre à la problématique scientifique énoncée et aux deux questions de recherche, ce travail de thèse s'est articulé autour de deux axes principaux. D'une part, nous avons conduit un travail de modélisation individuelle, multi-factorielle et temporelle du phénomène de polarisation. D'autre part, nos travaux ont porté sur le développement d'une approche personnalisée d'apport en diversité favorisant le contrôle de l'impact des recommandations sur les utilisateurs. Nous avons fait le choix de valoriser ces travaux à plusieurs reprises, et en particulier dans la conférence de référence dans le domaine de la modélisation utilisateur, et donc au plus proche de nos intérêts de recherche, la conférence UMAP (*User Modeling, Adaptation and Personalization*). Le détail de chacune de ces contributions est donné dans les sous-sections suivantes.

2.1 Modélisation multi-factorielle des comportements de polarisation

De nombreux chercheurs, issus de plusieurs domaines, ont consacré leurs travaux à la modélisation de la polarisation [Sîrbu *et al.*, 2017]. Cependant, selon nous, malgré la proposition de modèles avancés, la distinction entre les utilisateurs polarisés et non polarisés demeure complexe à établir. La principale limite de ces modèles réside dans la considération d'un unique facteur, ils ne parviennent donc pas à appréhender pleinement la complexité des comportements de polarisation qui sont influencés par une variété de facteurs tels que les sources d'information [Tokita *et al.*, 2021], les émotions exprimées [Buder *et al.*, 2021], le caractère controversé des débats [Garimella *et al.*, 2021], *etc.* En ne considérant qu'un unique facteur, les modèles peinent donc à fournir une modélisation fine du phénomène de polarisation.

Cette thèse contribue à cette modélisation en proposant une approche multi-factorielle pour modéliser les comportements de polarisation. A la fois informée par des concepts des sciences sociales et reposant sur l'exploitation de concepts mathématiques, cette approche de modélisation intègre ainsi la complexité inhérente au phénomène de polarisation. En appliquant l'approche proposée sur des données d'interactions et en appliquant des approches d'apprentissage machine comme le *clustering*, il devient possible de distinguer différentes classes de comportements de polarisation, jusqu'alors non identifiées. Ces classes représentent un grand intérêt dans une perspective de réduction de la polarisation.
Ce travail a fait l'objet de deux publications en conférence internationale :

— **C. Treuillier**, E. Dufraisse, S. Castagnos & A. Brun, (2022) Being Diverse is Not Enough : Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems, *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP'22)*

— **C. Treuillier**, S. Castagnos & A. Brun, (2023) A Multi-Factorial Analysis of Polarization on Social Media, *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP'23)*

2.2 Mesure individuelle et multi-factorielle de polarisation : GRAIL

En complément de la modélisation de la polarisation, plusieurs études se sont intéressées à sa mesure et à sa quantification [Esteban and Ray, 1994]. Cependant, les métriques de polarisation proposées sont généralement globales, permettant d'évaluer la polarisation à l'échelle d'un réseau, d'une communauté ou sur une thématique spécifique. Dans un contexte où les approches se veulent personnalisées, la quantification de la polarisation individuelle s'avère néanmoins plus pertinente. Malheureusement, ce type de métrique reste rare dans la littérature, et les quelques exceptions présentent une limite majeure : elles ne rendent pas compte de l'ensemble des comportements adoptés par un utilisateur particulier dont on tente de quantifier la polarisation.

Pour pallier ce manque et en prolongement du travail sur la modélisation multi-factorielle de la polarisation, une métrique individuelle de polarisation a été développée, baptisée **GRAIL** (*GeneRalized AddItive poLarization*). Cette métrique repose sur l'utilisation d'un modèle mathématique appelé modèle additif généralisé [Hastie, 2017], permettant de combiner différents facteurs de polarisation au sein d'une même mesure. Contrairement aux principales métriques existantes, GRAIL permet de quantifier le niveau de polarisation de chaque utilisateur individuellement. De plus, elle a été conçue pour être applicable dans divers contextes et propose plusieurs paramètres optimisables.
Ce travail a fait l'objet d'une publication en conférence internationale :

— **C. Treuillier**, S. Castagnos & A. Brun, (2024) All Polarized but Still Different: a Multi-factorial Metric to Discriminate between Polarization Behaviors on Social Media, *39th ACM/SIGAPP Symposium On Applied Computing (SAC'24)*

2.3 Modélisation temporelle des comportements de polarisation

La modélisation de la polarisation en tant que phénomène statique, bien qu'enrichissante, offre une perspective limitée puisqu'elle ne prend pas compte de l'évolution temporelle. En effet, la polarisation est un processus dynamique, influencé par des événements contextuels, des changements sociétaux et des interactions entre individus ou groupes [Baldassarri and Bearman, 2007]. La prise en compte de ces facteurs temporels est essentielle pour comprendre la nature évolutive de la polarisation. Une modélisation capable d'exploiter la dimension temporelle des données permettrait ainsi d'identifier des périodes de changement significatif, de reconnaître des tendances dans les comportements et de prédire les évolutions futures. Cela pourrait également aider à élaborer des stratégies d'intervention plus efficaces pour atténuer la polarisation ou pour promouvoir le dialogue et pacifier les débats.

Une contribution supplémentaire de cette thèse, qui fait suite aux précédentes et les complète, réside en une modélisation temporelle de la polarisation. Elle a notamment permis de définir une suite de périodes clés de polarisation, chacune étant caractérisée par la présence de comportements de polarisation spécifiques. L'intégration de cette dimension temporelle dans la modélisation de la polarisation enrichit la compréhension du phénomène. Elle ouvre la voie à des analyses plus nuancées et à des interventions plus ciblées pour contrôler la polarisation.
Ce travail a fait l'objet d'une publication en journal international :

— **C. Treuillier**, S. Castagnos, C.Lagier & A.Brun (2024) Gaining a better understanding of online polarization by approaching it as a dynamic process, *Scientific Reports*

2.4 Approche de diversification des recommandations contrainte par l'équité : ADF

Faisant suite à l'approche de modélisation, mon travail s'est davantage porté sur le processus de recommandation en lui-même. Comme détaillé dans le début de cette introduction, les recommandations de news doivent être adaptées de façon à exposer les utilisateurs à un contenu plus diversifié afin d'éviter le renforcement de la polarisation [Bernstein *et al.*, 2020]. La littérature regorge d'approches de diversification, qui cherchent à optimiser un compromis entre la diversité et la précision des résultats. Cependant, cette diversification se fait souvent de façon peu contrôlée, favorisant l'apparition de nouveaux biais comme la sur-représentation de certains types de contenu ou l'orientation artificielle des intérêts des utilisateurs vers des contenus très éloignés de leurs préférences initiales. Ceci soulève notamment des limites éthiques, en particulier en lien avec l'équité des recommandations diversifiées.

C'est dans ce contexte que s'inscrit la quatrième contribution : une approche post-traitement de diversification des recommandations appelée **ADF** *(Accuracy, Diversity, Fairness)*. Cette dernière repose sur une approche de diversification contrainte par l'équité, permettant d'assurer des recommandations à la fois équitables et diversifiées, tout en maintenant un niveau de précision acceptable. Ce qui la distingue de la littérature, c'est sa capacité à contrôler simultanément les trois dimensions essentielles des systèmes de recommandation que sont la pertinence, la diversité et l'équité.

Ce travail a fait l'objet d'une publication dans une conférence internationale :

— **C. Treuillier**, S. Castagnos, Ö. Özgöbek & A. Brun, (2024) Beyond Trade-offs : Unveiling Fairness-Constrained Diversity in News Recommender Systems, *Proceedings of the 32st ACM Conference on User Modeling, Adaptation and Personalization (UMAP'24)*

3 Un travail de thèse riche en collaborations

La collaboration est souvent la clé de la réussite dans la recherche académique, et le travail présenté dans ce manuscrit en est un bon exemple. En premier lieu, il est d'abord le résultat d'un travail conjoint réalisé avec mes deux directeurs de thèse, Armelle Brun (Professeure à l'Université de Lorraine, responsable scientifique de l'équipe BIRD (*Building artificial Intelligence Between Trust, Responsability and Decision*)) et Sylvain Castagnos (Maître de conférences à l'Université de Lorraine). Bien que ce manuscrit soit ensuite rédigé à la première personne, l'ensemble des travaux présentés sont le résultat de notre travail commun durant les trois années de thèse.

Par ailleurs, cette thèse s'inscrit dans le cadre d'un projet financé par l'Agence Nationale de Recherche (ANR), le projet BOOM[3] (*Modeling and Opening Opinion Bubbles*). L'objectif principal est de contribuer à la dépolarisation en proposant des solutions résultant d'une expertise pluridisciplinaire (sciences politiques, économie numérique, traitement automatique des langues et systèmes de recommandation). Au sein de ce consortium, j'ai notamment collaboré avec Evan Dufraisse, Julien Tourille et Adrian Popescu, attachés au laboratoire LIST du CEA[4], travaillant sur les aspects de traitement automatique des langues. Ensemble, nous avons pu mettre à profit les compétences et connaissances de nos domaines respectifs, ce qui a notamment permis une réflexion sur le besoin de collaborations pluri-disciplinaires pour répondre à des problématiques comme la polarisation [Dufraisse *et al.*, 2022], ainsi qu'un travail en cours sur une diversification ciblée sur les sentiments exprimés. J'ai également eu l'opportunité de travailler avec Christèle

3. http://boom.loria.fr/
4. https://list.cea.fr/fr/

Lagier (Maîtresse de conférences à l'Université d'Avignon et directrice du laboratoire JPEG [5] (*Laboratoire des sciences Juridiques, Politique, Économiques et de Gestion*)). Cette collaboration avec une spécialiste des sciences politiques a permis d'apporter une vision différente quant à la modélisation de la polarisation. Ce travail a permis de mettre en perspective les points de vue des sciences politiques, de l'apprentissage automatique et de la modélisation utilisateur. Les travaux portant sur les dynamiques de polarisation, et donc sur la modélisation temporelle sont le résultat de cette collaboration [Treuillier *et al.*, 2024a]. Évidemment, tous les membres du projet ont participé, de près ou de loin, à la réalisation et à la réussite des différents travaux entrepris.

Finalement, j'ai eu l'opportunité de profiter du dispositif DrEAM [6] proposé par l'Université de Lorraine. Ce dispositif soutient la mobilité internationale, en participant au financement d'un séjour de recherche des doctorants sélectionnés. J'ai ainsi passé 4 mois à l'Université Norvégienne des Sciences et Technologies (NTNU) de Trondheim en fin d'année 2023. J'ai rejoint l'unité DART [7] (données et intelligence artificielle) du département d'informatique, où j'ai collaboré avec Özlem Özgöbek (Maîtresse de conférences à NTNU) sur les aspects en lien avec la recommandation. Le framework ADF est un premier résultat de ce travail commun [Treuillier *et al.*, 2024b], et les collaborations se poursuivent actuellement.

4 Contenu du manuscrit

Dans l'objectif de répondre aux questions de recherche posées, la suite de ce manuscrit s'est naturellement organisée en deux parties distinctes. La première aborde la modélisation multi-factorielle et temporelle des dynamiques de polarisation et répond donc à la première question de recherche (*QR1*). Le Chapitre 1 débute par un état de l'art détaillé sur le phénomène de polarisation en ligne, permettant d'introduire les notions clés et de donner un aperçu des travaux existants sur la modélisation de la polarisation. Les contributions propres à cette thèse sont ensuite détaillées dans les chapitres suivants : la modélisation multi-factorielle de la polarisation et la métrique GRAIL sont présentées dans le Chapitre 2, tandis que la modélisation temporelle est présentée dans le Chapitre 3.

Le manuscrit se poursuit avec la seconde partie, traitant de l'apport en diversité dans les recommandations de news, et répondant donc à la seconde question de recherche (*QR2*). De la même façon que précédemment, cette partie débute avec un état de l'art sur les systèmes de recommandation et leur adaptation au contexte des news en Chapitre 4. La contribution sur cette seconde partie est détaillée à la suite, avec la présentation du framework ADF en Chapitre 5.

Bien que le manuscrit soit organisé en deux parties distinctes, les contributions associées n'en restent pas moins liées puisque la première partie sur la modélisation informe la seconde sur la recommandation. Une conclusion sur l'ensemble des travaux est donnée à la fin du manuscrit. Cette dernière permet d'étayer les différentes réflexions produites à la suite de ce travail de thèse, et offre de nouvelles perspectives de recherche.

5. https://univ-avignon.fr/laboratoire/laboratoires-en-sciences-humaines-et-sociales/upr-3788-jpeg-laboratoire-sciences-juridiques-politique-economiques-gestion/
6. https://doctorat.univ-lorraine.fr/fr/international/dream
7. https://www.ntnu.edu/idi/dart#/view/about

Première partie

Modélisation multi-factorielle des dynamiques de polarisation

Chapitre 1

État de l'art : le phénomène de polarisation en ligne

Sommaire

Cette première partie du manuscrit s'ouvre par un état de l'art portant sur le phénomène de polarisation, au cœur de ma problématique de thèse. La polarisation est un phénomène complexe étudié par des chercheurs de tous horizons. Après une présentation générale, cette revue de la littérature se focalisera tout d'abord sur le rôle des médias sociaux dans le perpétuation du phénomène social de polarisation. Ensuite, divers travaux de modélisation informatique et mathématique seront détaillés, puisque directement en lien avec ma première question de recherche. Une section de conclusion permettra de faire le point sur l'état actuel de la littérature sur le sujet, permettant ainsi par le même temps de positionner le travail qui est présenté dans la suite de la partie.

1.1 La polarisation : un phénomène complexe

1.1.1 Définition de la polarisation

La polarisation, dans son sens le plus large, désigne l'"attraction vers ou autour d'un ou plusieurs pôle(s), sujet(s), thème(s)" [8]. Initialement, le terme de polarisation était attaché à des concepts physiques ou mathématiques, avec notamment le concept de polarisation des ondes. Le terme a ensuite évolué pour englober des significations plus abstraites, en particulier dans le domaine des sciences sociales, où la polarisation peut alors être définie comme la "concentration de l'attention, des activités, des influences, des sentiments sur un même point" [7]. Le terme polarisation, tel qu'il est employé usuellement dans les médias, dans la littérature scientifique et dans la suite de ce manuscrit, fait en réalité référence au phénomène de *polarisation politique*, définie ci-dessous.

> ***Polarisation politique***
> État de division profonde au sein d'une société, où les opinions et les attitudes politiques des citoyens tendent à se regrouper, en diminuant le terrain d'entente.

Cette définition de polarisation politique n'est pas unique, et il n'en n'existe pas de consensus. Cependant, bien que les définitions soient multiples, une caractéristique est partagée : elle se manifeste principalement par une concentration des opinions au sein de groupes, avec la création de ***liens intra-communautaires***, avec un rejet des idées extérieures au groupe, donc une augmentation de la ***distance inter-communautaire*** [Sunstein, 2009]. Cet attachement à des groupes et ce rejet de points de vue externes est le résultat de multiples facteurs, ce qui fait de la polarisation un phénomène complexe [Levin *et al.*, 2021]. Elle peut entraîner l'adoption de comportements extrêmes, illustrés par des événements tels que l'assaut du Capitole des États-Unis par des partisans de Donald Trump le 6 janvier 2021. Les États-Unis sont d'ailleurs un très bon exemple de polarisation, largement abordé dans la littérature de par sa structuration bipartite, affrontant démocrates et républicains : les partisans sont très proches des idées exprimées par leur parti, et sont extrêmement imperméables aux idées du parti adverse avec lequel ils sont irréconciliables [Fiorina and Abrams, 2008]. Évidemment, la polarisation s'étend bien au-delà des frontières américaines et opère dans le monde entier [Carothers and O'Donohue, 2019].

Perspectives des sciences cognitives

Du point de vue des sciences cognitives, la polarisation implique une série de mécanismes cognitifs qui modulent la façon dont les individus traitent l'information et interagissent avec leur environnement politique. En particulier, le concept de dissonance cognitive [Festinger, 1962] survient lorsque les individus ressentent un inconfort psychologique lorsqu'ils sont confrontés à des informations, des croyances ou des attitudes contradictoires aux leurs. Pour atténuer cette dissonance, ils ont alors tendance à ajuster leurs croyances, attitudes et comportements. Cela peut se traduire par une recherche sélective d'informations conformes à leurs convictions actuelles ou par une interprétation sélective des informations qui renforcent leurs convictions préexistantes, respectivement appelées exposition sélective et perception sélective [Perrissol and Somat, 2009]. Ces divers mécanismes cognitifs contribuent à la formation de biais de confirmation [Klayman, 1995], qui désignent la propension naturelle à chercher des informations confirmant les croyances ou hypothèses préexistantes. La polarisation est donc médiée par des processus cognitifs, communs à l'ensemble des individus.

8. https://www.cnrtl.fr/lexicographie/polarisation

Perspectives des sciences sociales et politiques

Le processus de polarisation s'inscrit également dans un contexte social, politique et informationnel dynamique et complexe, influencé par de nombreux autres facteurs.

En sciences sociales, l'homophilie décrit la propension naturelle des individus à fréquenter des personnes semblables, partageant des caractéristiques communes [McPherson *et al.*, 2001]. En partageant préférentiellement des opinions similaires aux leurs, les individus ne se confrontent pas à des opinions divergentes et renforcent leur polarisation [Dandekar *et al.*, 2013].

En sciences politiques, de multiples phénomènes sociaux sont étudiés depuis de nombreuses années et permettent d'expliquer l'existence du phénomène de polarisation, comme la répartition inégale des compétences politiques [Carpini and Keeter, 1996], l'évitement de la politique [Eliasoph, 1998] ou encore les effets contextuels [Huckfeldt, 1986] qui jouent un rôle crucial dans la formation et le renforcement des attitudes polarisées.

1.1.2 Impact sociétal de la polarisation et enjeux associés

Malgré la multiplication des études sur la polarisation, son réel impact au niveau sociétal reste également largement discuté [Heltzel and Laurin, 2020]. Bien que la notion de polarisation soit majoritairement associée à des conséquences négatives, certains chercheurs soulignent son intérêt. L'existence de tension permet, d'une part, d'entretenir la dimension démocratique d'une société, au sein de laquelle chacun est libre d'exprimer son point de vue [Abrams, 2007]. Cette polarisation représente, d'autre part, l'opportunité pour les citoyens de s'engager dans le débat public [Kriesi, 2017]. Au sein de la population, les individus polarisés sont donc avant tout ceux qui ont la capacité de formuler une opinion structurée sur les débats de société, et plus cette opinion est affirmée, moins elle est susceptible d'être modifiée [Katz *et al.*, 2017]. Ces individus ne sont pas simplement victimes d'un filtrage personnalisé de l'information, de la circulation de fausses informations ou de phénomènes de contagion morale [Goldstein, 1984], comme évoqué dans de nombreuses analyses de la dynamique d'opinion [Sîrbu *et al.*, 2017]. Ainsi, en examinant plus précisément les dynamiques de polarisation, il devient possible d'élaborer des stratégies pour promouvoir le dialogue constructif et réduire les tensions, tout en préservant la diversité des points de vue et l'engagement démocratique. De mon point de vue, l'objectif consiste ainsi à favoriser une compréhension plus approfondie de la polarisation dans le contexte de l'information en ligne. C'est selon moi un pré-requis précieux pour développer des approches plus efficaces dans la gestion des divergences d'opinions et des dérives qui y sont associées.

Pour répondre à un tel objectif, il est primordial de rendre compte de la complexité du phénomène de polarisation, qui est influencé par une multitude de facteurs et exacerbé par l'évolution des sociétés et des technologies de l'information. Du point de vue des sciences sociales, la polarisation prend différentes formes, avec notamment une distinction entre la *polarisation idéologique*, où les extrêmes politiques gagnent en popularité au détriment des positions centristes, et la *polarisation affective*, qui souligne l'aspect émotionnel des clivages politiques, où les individus éprouvent des sentiments hostiles envers ceux de l'autre bord politique [Iyengar *et al.*, 2019]. Cette dynamique n'est pas limitée aux élites politiques [Hetherington, 2001] mais s'étend à la société dans son ensemble, créant une polarisation de masse [Abramowitz and Saunders, 2008] qui reflète un électorat de plus en plus fragmenté le long de lignes idéologiques.

Il convient également de souligner que la polarisation ne se limite pas à l'espace public ou aux interactions en face à face : elle se propage également dans l'environnement numérique, engendrant une polarisation en ligne [Waller and Anderson, 2021]. Dans ce contexte, il apparaît crucial d'analyser l'impact des outils numériques sur la polarisation politique et de comprendre

comment ils façonnent les débats publics. C'est précisément ce qui est au cœur du travail que je présente dans cette première partie du manuscrit, et qui est abordé au travers de la première question de recherche (*QR1*). La recherche dans ce domaine peut en effet offrir des indicateurs précieux pour concevoir ensuite des stratégies visant à atténuer les effets négatifs de la polarisation. La section suivante présente les recherches menées en ce sens, en mettant l'accent sur le rôle des algorithmes et des outils numériques dans le processus de polarisation.

1.2 La polarisation en ligne : rôle des médias sociaux

Depuis la fin des années 2000, la démocratisation du numérique a considérablement modifié les pratiques informationnelles des citoyens [Patino, 2019]. L'information est disponible gratuitement, facilement et en grande quantité, laissant l'opportunité à tous et à tout moment de s'informer sur des sujets variés [Cagé *et al.*, 2020]. Les médias sociaux regroupent à la fois les médias en ligne, par exemple la version numérique de journaux originellement papier tel que Le Monde [9], les agrégateurs d'actualité comme Google Actualités [10], et les réseaux sociaux comme Facebook [11], X [12], Instagram [13] et bien d'autres. Ces nouveaux écosystèmes médiatiques participent activement à renforcer la polarisation [Kubin and Von Sikorski, 2021, Van Bavel *et al.*, 2021]. D'une part, la propagation de fausses informations (*fake news*) est largement facilitée et accélérée par ces médias [Lazer *et al.*, 2018]. Ces informations trompeuses influencent et manipulent l'opinion publique, participant au phénomène de désinformation pouvant renforcer l'adoption d'opinions extrêmes et alimentant la polarisation [Azzimonti and Fernandes, 2023]. D'autre part, les mécanismes de filtrage de l'information mis en place sur les différentes plateformes servent aussi de catalyseurs de la polarisation [Zuiderveen Borgesius *et al.*, 2016]. En effet, bien que conçus pour aider les utilisateurs à trouver le contenu d'intérêt parmi la large quantité de données disponible, ces filtres participent à réduire la diversité des contenus à laquelle les utilisateurs sont confrontés, exacerbant ainsi la polarisation. C'est ce deuxième point sur lequel je me focalise, et qui est détaillé dans les sous-sections suivantes.

1.2.1 Le filtrage de l'information dans les médias et réseaux sociaux

Comme mentionné ci-dessus, l'abondance de données permise par Internet et les médias sociaux a rendu le filtrage de l'information indispensable. Ce processus de sélection est essentiel pour aider les utilisateurs à identifier les contenus d'intérêt parmi la vaste quantité de données disponible. Il doit être adapté aux besoin de chacun des utilisateurs, et est donc personnalisé. Grâce à ce processus, le contenu médiatique ou informationnel qui est consulté par les utilisateurs n'est pas le même pour tous, mais adapté à différents groupes ou individus. Deux types de personnalisation doivent être distinguées [Zuiderveen Borgesius *et al.*, 2016] :

— **La personnalisation auto-sélectionnée (personalisation explicite) :** les individus choisissent par eux-mêmes de s'informer exclusivement avec certaines sources d'information, partageant les mêmes opinions, et tendent ainsi à éviter les informations mettant en cause leurs croyances initiales. Cette personnalisation joue un rôle crucial dans la formation de communautés d'opinions. Ces communautés, souvent appelées *chambres d'écho*, sont des espaces où les idées et les croyances sont renforcées par la répétition au sein d'un groupe

9. https://www.lemonde.fr/
10. https://news.google.com/
11. https://www.facebook.com/
12. https://twitter.com/
13. https://www.instagram.com/

homogène [Sunstein, 1999]. Ce phénomène peut conduire à une polarisation des opinions et à une diminution de l'exposition à des points de vue diversifiés, ce qui peut effectivement altérer la perception de la réalité des individus. Il est donc essentiel de promouvoir un échange d'idées plus ouvert et diversifié pour maintenir un équilibre sain dans le paysage informationnel.

Chambre d'écho
Communauté d'individus partageant des opinions similaires, rejetant activement les idées contraires, et où la perception de la réalité est faussée.

— **La personnalisation pré-sélectionnée (personnalisation implicite) :** résultat d'un filtrage automatisé du contenu présenté aux utilisateurs. Dans ce cas, il y a souvent un manque de transparence car les utilisateurs ne sont pas tous conscients de la présence de tels systèmes. Par ailleurs, pour filtrer le contenu, ces algorithmes exploitent les historiques de consommation des utilisateurs afin d'identifier le contenu le plus susceptible de répondre à leurs attentes. Bien qu'améliorant leur expérience et favorisant l'acceptation des systèmes, ce mécanisme entraîne la consommation de contenus peu diversifiés [Bozdag, 2013]. Cette exposition sélective limite la présentation de points de vue contrastés. Les utilisateurs se retrouvent alors piégés dans des bulles de filtre [Pariser, 2011, Bozdag and van den Hoven, 2015].

Bulle de filtre
État d'isolement intellectuel ou idéologique résultant d'un filtrage automatisé de l'information.

Bien que les concepts de chambres d'écho et de bulles de filtre soient largement abordés dans la littérature, ils sont souvent confondus. Ce sont pourtant deux choses bien différentes. Les bulles de filtre sont le résultat d'algorithmes de filtrage seuls, tandis que les chambres d'écho peuvent être le résultat d'autres processus de filtrage, parfois volontaires de la part des individus (Figure 1.1). Le psychologue C. Thi Nguyen résume cette différence en expliquant que les bulles de filtre sont des espaces où les utilisateurs n'*entendent* pas les autres points de vue, tandis qu'au sein de chambres d'échos, ils ne *croient* pas aux autres points de vue [Nguyen, 2020].

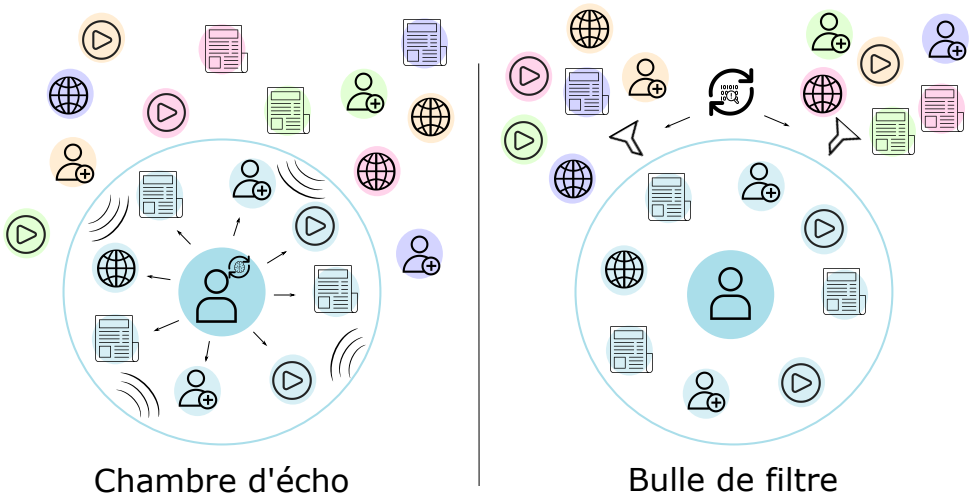


FIGURE 1.1 – Chambre d'écho *vs.* Bulle de filtre

Néanmoins, ce qui est commun aux chambres d'écho et aux bulles de filtre est l'isolement informationnel des individus, lié à une réduction de la diversité des informations consommées. Cela favorise la polarisation des opinions [Zuiderveen Borgesius *et al.*, 2016]. Sur le long terme et lorsqu'elle atteint un niveau trop extrême, la polarisation entraîne l'adoption de comportements extrêmes ou violents [Sunstein, 2009] et devient alors préjudiciable au débat politique et, en fin de compte, aux sociétés démocratiques. Les médias sociaux jouent donc un rôle central dans le phénomène de polarisation, et peuvent avoir d'importantes conséquences.

1.2.2 Impact des médias sociaux sur le phénomène de polarisation

L'étude du rôle et de l'impact des médias sociaux sur la polarisation a gagné l'intérêt des chercheurs depuis quelques années, avec notamment une importante évolution du nombre de papiers rédigés sur le sujet depuis 2012 [Kubin and Von Sikorski, 2021]. Il est cependant important de noter que les conclusions tirées sur le sujet il y a une dizaine d'années peuvent être obsolètes car l'utilisation des médias sociaux, la nature des contenus publiés, l'animosité entre les opposants politiques et donc la polarisation elle-même ont largement évolué. De plus, la plupart des études s'intéressent au contexte américain, où le système bipartite domine le paysage politique. Cependant, les médias sociaux sont utilisés à travers le monde et les comportements de polarisation sont impactés quel que soit le pays et son système politique. Des études ont par exemple été menées au Chili [Valenzuela *et al.*, 2021], au Guana [Conroy-Krutz and Moehler, 2015], en Allemagne [Knobloch-Westerwick *et al.*, 2015] et en France [Peralta *et al.*, 2024]. Les conclusions tirées dans le contexte américain sont donc à considérer avec précaution car ne sont pas nécessairement applicables dans d'autres contextes, où le fonctionnement du système politique est différent.

Polarisation des contenus et exposition aux médias sociaux

Pour évaluer l'influence des médias sociaux sur la polarisation, il est possible de distinguer l'impact des contenus diffusés de celui de l'exposition à ces médias [Kubin and Von Sikorski, 2021].

Dans un premier temps, les contenus diffusés peuvent eux-mêmes être polarisés ou polarisant. Cette polarisation des contenus opère à la fois sur les réseaux sociaux, il a par exemple été montré que les contenus devenaient plus polarisés au cours du temps sur le réseau social Twitter [Marozzo and Bessi, 2018], et sur certains médias traditionnels [Hyun and Moon, 2016]. Ainsi, les médias très spécifiques, partisans de partis politiques extrêmes adoptent parfois des rhétoriques chocs, polarisantes, de façon à frapper leurs lecteurs et favorisent ainsi l'adoption ou le renforcement de positions extrêmes. De la même façon, certains politiciens qui utilisent les réseaux sociaux comme moyen de communication de leurs idées, emploient des discours précis de façon à répandre leurs messages [Russell, 2021].

Dans un second temps, l'impact du filtrage de l'information et de l'exposition sélective sur les médias sociaux a été largement étudié. Un lien entre l'exposition aux médias et l'augmentation de la polarisation a notamment été affirmé [Van Stekelenburg, 2014]. [Chang and Park, 2021] suggèrent même une relation réciproque entre l'exposition aux médias et l'augmentation de la polarisation. Ces mêmes chercheurs mettent également en avant que l'utilisation des médias sociaux est liée à une augmentation de la participation aux protestations politiques. Les utilisateurs susceptibles de se polariser en ligne sont donc aussi ceux qui ont la capacité à se politiser, et donc à s'engager dans un processus politique. Cet engagement est particulièrement bien représenté sur le réseau social Twitter, dont les utilisateurs ont une sensibilisation accrue aux questions politiques [Boyadjian *et al.*, 2014]. Ils y ont l'opportunité d'y partager, en de très courts textes, leurs

opinions. Par ailleurs, en utilisant notamment les *retweets* pour partager le contenu d'un utilisateur ou les mentions pour le partager en y ajoutant un commentaire, les utilisateurs peuvent aisément exprimer leur accord ou désaccord avec les informations diffusées. Cinq ans après la publication du premier tweet en 2006, Conover et al. [Conover *et al.*, 2011] ont d'ailleurs montré que ces interactions (mentions et *retweets*) induisent des topologies de réseau distinctes. Le réseau de *retweets* permet de distinguer deux communautés bien séparées, montrant ainsi qu'il existe une forte polarisation, tandis que le réseau de *mentions* donne lieu à une communauté unique très connectée. Cela confirme à la fois l'existence d'une forte polarisation sur le réseau social et démontre également que les divers types d'interactions influencent les comportements et les dynamiques d'opinions de manière distincte. Par ailleurs, de nombreuses études portent également sur le réseau social Facebook, qui reste encore aujourd'hui le réseau social le plus utilisé à travers le monde, avec plus d'un milliard de comptes créés. En particulier, [Del Vicario *et al.*, 2016] confirment que Facebook est un environnement particulièrement adapté à l'émergence de communautés polarisées, ou chambres d'écho. Leurs résultats montrent également que les utilisateurs les plus actifs sont ceux qui sont les plus susceptibles de dévier rapidement vers des comportements extrêmes. Par exemple, sur le débat spécifique de la vaccination, les utilisateurs consomment préférentiellement du contenu au sein de leur communauté, ce qui renforce la polarisation au cours du temps sur cette thématique [Schmidt *et al.*, 2018]. Finalement, [Bessi *et al.*, 2016] confirment à leur tour l'existence de chambres d'écho sur Facebook, et une analyse comparative permet de mettre en avant la formation de communautés similaires sur Youtube [14]. Le phénomène de polarisation est donc susceptible d'être renforcé sur une large variété de réseaux sociaux, allant des plus communément utilisés comme Facebook, Twitter et TikTok [15] [Lund and Zhong, 2018], aux plus inattendus comme Youtube, en passant par des réseaux sociaux moins populaires comme Reddit [16] [Jasser *et al.*, 2022].

Remise en question de l'impact des médias sociaux

L'impact de l'utilisation des médias sociaux sur le renforcement du phénomène de polarisation reste contesté [Prior, 2013]. En ce sens, Axel Bruns remet en question l'existence même des bulles de filtre, et expose notamment un manque de preuves sur les effets associés [Bruns, 2019]. Certaines études expliquent également que l'exposition aux médias permet de se confronter à une large diversité d'informations [Fletcher and Nielsen, 2018], et donc que les utilisateurs de ces médias sociaux seraient exposés à une plus large diversité d'information que les individus qui ne les utilisent pas. Selon [Haim *et al.*, 2018], le filtrage de l'information sur les médias sociaux, et plus particulièrement sur l'agrégateur Google Actualités, n'entraînerait pas de sur-personnalisation de l'information. [Beam *et al.*, 2020] avancent des conclusions similaires, en démontrant que les utilisateurs de Facebook sont plus susceptibles d'être confrontés à du contenu diversifié, allant dans leur sens (pro-attitudinal) ou non (contre-attitudinal). L'impact de l'utilisation des médias sociaux et du filtrage de l'information, qui est donc largement débattu dans la littérature, peut justement se différencier en fonction de la nature du contenu auquel les utilisateurs sont exposés. Dans un premier cas, les conclusions tirées sont plus cohérentes lorsqu'il s'agit d'une exposition à du contenu pro-attitudinal, renforçant le phénomène de polarisation [Knobloch-Westerwick *et al.*, 2015, Kim, 2015]. Au contraire, lorsque les utilisateurs sont confrontés à du contenu contre-attitudinal, les effets sont mixtes : bien que promouvant la prise de connaissance par l'exposition a un contenu plus divers, le contenu dissonant peut renforcer

14. https://www.youtube.com/
15. https://www.tiktok.com/
16. https://www.reddit.com/

la polarisation dans certains cas [Kim, 2019]. Ce renfort de la polarisation par l'exposition à un contenu contraire aux croyances, et faisant donc écho au phénomène de dissonance cognitive [Festinger, 1962] détaillé au début de cet état de l'art, a été mis en évidence à plusieurs reprises. [Bail *et al.*, 2018] ont par exemple montré que les républicains, au contraire des démocrates, ont cette tendance à renforcer leur position lors de l'exposition aux arguments du camp adverse. Évaluant à la fois le contexte américain et israëlien, dont le fonctionnement politique est différent, [Garrett *et al.*, 2014] tirent des conclusions similaires, en mettant en avant le potentiel polarisant de l'exposition à un contenu contradictoire. Bien que l'exposition à une plus large variété de contenus semble être une approche intuitive pour aider à réduire la polarisation, elle peut donc en réalité la renforcer. Ainsi, de mon point de vue, les approches potentielles visant à atténuer la polarisation doivent être manipulées avec précaution afin d'en garantir l'effet positif sur la façon dont les utilisateurs consomment l'information.

Pour conclure, bien que certaines études remettent en question le véritable rôle des réseaux sociaux et de l'exposition sélective sur le phénomène de polarisation [Bruns, 2019], le lien entre la consommation de l'actualité en ligne et ce phénomène démocratique de notre société actuelle n'en reste pas moins évident [Levy, 2021]. Dans ce contexte, l'absence d'une définition commune de la polarisation rend difficile l'évaluation de l'impact global des médias sociaux. Les résultats variables d'une étude à l'autre sont également probablement dus aux différences de conceptualisation de la polarisation [Barberá, 2020]. Les conclusions tirées sur la polarisation sont le résultat d'études très hétérogènes, appliquées à des contextes divers. Pour évaluer l'impact des réseaux sociaux de façon plus fine, et donc pour répondre à la problématique guidant ce travail de thèse, une modélisation adaptée du processus de polarisation m'apparaît nécessaire.

1.3 Modélisation de la polarisation

La section précédente a mis en lumière la complexité de la polarisation de par la multiplicité des facteurs associés, qui peuvent être de nature cognitive (biais de confirmation), sociale (homophilie), politique (politisation) ou technologique (filtrage automatisé de l'information). Une compréhension approfondie de la polarisation nécessite donc des connaissances provenant de diverses disciplines [Jung *et al.*, 2019]. En plus de connaître la nature des différents facteurs sous-jacents, **une représentation mathématique et informatique des dynamiques qui sous-tendent la polarisation aiderait à mieux la caractériser et mieux la comprendre**. Les modèles de polarisation vont des approches basées sur des équations mathématiques [Sîrbu *et al.*, 2017], aux simulations informatiques [Baumann *et al.*, 2020], en passant par les modèles statistiques [Kaufman *et al.*, 2022]. Ils tentent souvent de reproduire des phénomènes observés dans des contextes réels ou simulés, tels que la formation de bulles de filtre [Geschke *et al.*, 2019], la fragmentation des réseaux sociaux [Conover *et al.*, 2011], ou encore l'émergence de consensus ou de conflits [Garimella *et al.*, 2018].

La polarisation est principalement appréhendée au travers des dynamiques d'opinion qui y sont associées. Ces opinions guident les actions des individus et leur formation découle d'un processus complexe influencé par les informations auxquelles ils sont exposés [Sîrbu *et al.*, 2017]. La compréhension des dynamiques sous-jacentes peut ainsi permettre de gagner des connaissances quant au fonctionnement des comportements humains, comme ceux de polarisation.

1.3.1 Modèles originels de polarisation

Dans les années 70, le statisticien Morris DeGroot propose un modèle populaire pour la modélisation des dynamiques d'opinions [DeGroot, 1974]. Ce dernier permet de modéliser l'ajustement des opinions d'agents au cours du temps en fonction des opinions de leur entourage. A chaque itération, cette opinion est mise à jour en fonction des opinions voisines, qui sont pondérées en fonction de leur force de connexion. Ce modèle est souvent utilisé pour étudier la façon dont les individus parviennent à un consensus dans un groupe [Eger, 2016], mais est également utilisé pour étudier le processus de polarisation [Dandekar *et al.*, 2013]. En ce sens, le modèle DeGroot est parfois adapté pour tenir compte d'autres dimensions en lien avec le phénomène de polarisation. Par exemple, [Alvim *et al.*, 2023] modifient la pondération des opinions voisines lors de la mise à jour de celle d'un agent, de façon à modéliser les biais de confirmation.

Au delà des modèles statistiques inspirés du modèle DeGroot, des modèles reposant sur la modélisation multi-agents sont développés. Ce type de modélisation est particulièrement adapté à la tâche de modélisation de la polarisation en considérant plusieurs agents représentant les individus [Holland and Miller, 1991], qui peuvent communiquer entre eux et mettre à jour leur opinion en fonction de diverses règles [Miller and Page, 2008]. L'un des modèles les plus simples et les plus populaires, inspiré de la physique, est le modèle d'Ising [Baxter, 2016], qui représente l'opinion de chaque agent comme un état binaire, appelé *spin*, qui peut prendre uniquement deux positions ($+1$ ou -1). A chaque itération, cette opinion est mise à jour en fonction de celles des agents voisins selon certaines règles. L'opinion peut par exemple s'aligner sur l'opinion majoritaire des voisins. Ce modèle très simple, initialement décrit pour étudier les propriétés des matériaux magnétiques, a inspiré le développement de nombreux modèles plus aboutis dans des disciplines plus variées. Ces modèles se distinguent par le nombre de dimensions considérées (modèles unidimensionnels *vs.* modèles multi-dimensionnels), et la nature des opinions modélisées (discrètes *vs.* continues) [Sîrbu *et al.*, 2017].

1.3.2 Modèles enrichis de polarisation

Plus récemment, des modèles enrichis permettant de modéliser la polarisation sous l'influence de facteurs plus divers ont été proposés dans la littérature. Les très nombreux modèles existants ne pouvant être tous détaillés dans cet état de l'art, j'en ai sélectionné certains permettant de modéliser différents aspects de la polarisation.

Modélisation des dynamiques de polarisation

[Baumann *et al.*, 2020] ont par exemple modélisé les dynamiques de radicalisation, comme un renforcement des opinions extrêmes sur les réseaux sociaux. Pour cela, ils modélisent l'évolution des opinions x de N agents. Plus précisément, l'opinion x_i d'un agent i est influencée par celles des agents j voisins selon l'Équation (1.1).

$$\dot{x_i} = -x_i + K \sum_j A_{ij}(t) tanh(\alpha x_j) \tag{1.1}$$

où le paramètre K modélise la force de l'influence sociale entre les agents, le paramètre α contrôle le niveau de controversialité du sujet, et la fonction $tanh$ décrit l'influence de l'opinion d'un agent voisin j sur celle de l'agent i. Les voisins de l'agent i sont déterminés en fonction d'une matrice d'adjacence temporelle $A_{ij}(t)$, avec $A_{ij} = 1$ s'il existe un lien entre les agents i et j, et $A_{ij} = 0$ sinon. Pour aller plus loin, les auteurs ont raffiné ces interactions entre les agents

de façon à modéliser le phénomène d'homophilie en définissant la probabilité d'interaction entre un agent i et un agent j selon l'Équation (1.2).

$$p_{ij} = \frac{|x_i - x_j|^{-\beta}}{\sum_j |x_i - x_j|^{-\beta}} \tag{1.2}$$

où le paramètre β permet de contrôler la force du phénomène d'homophilie. Ce modèle permet donc de modéliser l'évolution de la polarisation en fonction de trois facteurs de polarisation : la controversialité du sujet (paramètre α), la force de l'influence sociale (paramètre K) et l'homophilie (paramètre β). En faisant varier ces différents paramètres, les auteurs sont capables d'étudier l'impact de chacun d'entre eux et de décrire l'évolution temporelle des agents. Trois types d'évolution temporelle des opinions sont mises en avant : le consensus, la polarisation unilatérale et la polarisation bilatérale. Cette dernière apparaît notamment lorsque les interactions concernent des sujets controversés (α élevé), pour lesquelles le mécanisme de renforcement social (K et β élevés) apparaît et favorise les dynamiques de polarisation. En 2021, ces mêmes auteurs proposent un modèle proche mais tenant compte de la multiplicité des sujets discutés sur les réseaux sociaux et de leur interaction [Baumann *et al.*, 2021]. Ils évoquent ainsi le concept d'alignement des enjeux (*issue alignment*), entraînant une corrélation des opinions. Les résultats mettent en avant de fortes corrélations entre les sujets lorsque ces derniers sont très controversés.

Les modèles proposés par Baumann et ses collaborateurs permettent donc de modéliser l'impact d'éléments sociaux (homophilie, influence sociale) ou liés au contenu des informations (controversialité des sujets, recouvrement des sujets) sur le niveau de polarisation d'un réseau. Suivant une approche similaire, [Macy *et al.*, 2021] proposent un modèle permettant d'étudier l'influence de la tolérance des agents aux idées contraires et la force de leur appartenance à leur parti sur le niveau de polarisation. Leurs résultats montrent que lorsque ces paramètres sont élevés, le modèle atteint un point de non retour (*tipping point*), au delà duquel une diminution de la polarisation est inenvisageable. L'application de chacun de ces modèles permet donc d'étudier l'influence de différents facteurs sur l'évolution de la polarisation au sein d'un réseau constitué d'agents réévaluant leurs opinions en fonction de leurs interactions.

Modélisation de l'impact du filtrage de l'information sur la polarisation

Comme expliqué au début de cet état de l'art, le filtrage de l'information joue un rôle important dans le renforcement de la polarisation en ligne. En 2019, [Perra and Rocha, 2019] proposent de modéliser l'effet de ce filtrage de l'information sur l'évolution de la polarisation. Ils proposent pour cela un modèle de formation d'opinions contenant trois parties : (1) un réseau social dont les connexions entre les utilisateurs sont établies en fonction de liens d'amitié ; (2) un mécanisme d'activation permettant de définir les moments où les informations sont échangées entre les utilisateurs du réseau social ; (3) un mécanisme de filtrage algorithmique qui sélectionne les informations présentées aux utilisateurs. Au sein de ce réseau, l'opinion de chaque utilisateur est évaluée au travers d'une probabilité $P(A > B)$, reflétant la probabilité que l'opinion de l'utilisateur soit plutôt en faveur de l'opinion A que de l'opinion B. Cette probabilité évolue en fonction des opinions (A et B) auquel l'utilisateur est confronté sous l'effet du filtrage de l'information. Les mécanismes de filtrages qui sont modélisés, évalués et comparés par les auteurs peuvent être aléatoires, temporels ou personnalisés et adaptés aux opinions de l'utilisateur. Les résultats permettent de mettre en évidence une relation entre le type de filtrage appliqué et l'évolution de la polarisation. Notamment, les auteurs montrent que lorsque les informations sont sélectionnées de façon à être proches des opinions des utilisateurs, la polarisation est renforcée.

Par ailleurs, [Valensise *et al.*, 2023] proposent également un modèle permettant de modéliser des facteurs algorithmiques, inspiré de celui de [Baumann *et al.*, 2020]. Ils introduisent notamment un paramètre permettant de consolider les liens existants et d'affaiblir les liens faibles, comme le ferait un algorithme de filtrage en recommandant du contenu proche des intérêts de l'utilisateur. Les résultats permettent de mettre en avant le rôle crucial de ce paramètre, qui a le pouvoir de contrôler le niveau de polarisation au sein du réseau d'agents. Ce modèle permet donc de modéliser plus précisément la polarisation en ligne, et confirme un impact important des algorithmes de filtrage sur les variations de niveau de polarisation.

Modélisation de l'impact de facteurs pluri-disciplinaires sur la polarisation

Une modélisation de la formation des chambres d'écho et des bulles de filtre, tenant à la fois compte d'aspects individuels, sociaux et technologiques, est proposée par [Geschke *et al.*, 2019]. Les auteurs proposent de représenter des agents dans un espace à 2 dimensions, au sein duquel ils évoluent temporellement en fonction de leur intégration des informations. Cette dernière est modélisée au travers d'un modèle probabiliste, présenté dans l'Équation (1.3).

$$P(d; D; \delta) = \frac{D^\delta}{d^\delta + D^\delta} \tag{1.3}$$

où d correspond à la distance entre l'information et l'utilisateur, D correspond à la latitude d'acceptation de l'utilisateur, *c.-à-d.* jusqu'à quelle distance l'utilisateur peut accepter et intégrer une information, et δ spécifie la pente avec laquelle la probabilité d'intégration change entre 0 et 1 autour de cette latitude d'acceptation. L'intégration d'une nouvelle information par un individu est donc modélisée à la fois en fonction de la distance de cette information et du niveau d'acceptation de l'utilisateur. Ceci reflète un premier facteur individuel de polarisation, lorsque le filtrage de l'information est opéré par l'individu lui-même, notamment sous l'effet de mécanismes cognitifs comme les biais de confirmation.

Au-delà de la probabilité d'intégration des informations, différents modes d'accès à ces informations sont modélisés. Les informations parvenant aux utilisateurs peuvent être individuelles, *c.-à-d.* chaque utilisateur reçoit une information qui lui est propre, centrales, *c.-à-d.* tous les utilisateurs reçoivent une même information, ou bien sélectionnées et être plus ou moins proches des utilisateurs, ce qui permet de modéliser l'effet d'un filtrage algorithmique de l'information. Cette seconde caractéristique du modèle permet donc de modéliser le facteur algorithmique de la polarisation, avec le filtrage automatisé de l'information.

Finalement, le modèle proposé permet d'établir des liens d'amitié entre les agents, qui peuvent ensuite partager des informations entre eux de façon préférentielle. Ce mécanisme reflète un facteur social de polarisation, avec le phénomène d'homophilie.

En faisant varier les paramètres liés à ces trois dimensions (individuelle, sociale et technologique) dans divers scénarios, les auteurs sont capables de refléter des phénomènes observés dans la vie réelle, notamment l'émergence de réseaux sociaux homophiles, de chambres d'écho et de bulles de filtre.

En conclusion, les modèles présentés reposent sur une multitude de facteurs susceptibles d'influencer le processus de polarisation, qu'il s'agisse de variables liées aux individus et à leurs comportements ou de l'effet du filtrage de l'information par les algorithmes. Les différents résultats permettent ainsi d'avoir une compréhension plus complète du phénomène, et de détailler l'impact des divers facteurs sur son évolution. Dans une perspective de réduction des effets négatifs de la polarisation, cette modélisation ouvre la voie à l'élaboration de stratégies adaptées.

1.4 Mesure de la polarisation

Outre la compréhension du phénomène de polarisation à travers sa modélisation, sa mesure revêt également une importance capitale car elle permet d'évaluer l'ampleur de la polarisation, d'identifier les tendances émergentes, de suivre l'efficacité des interventions visant à la réduire, *etc.* Dans cette section, les principales métriques de polarisation de la littérature sont présentées. La question de recherche posée dans le cadre de cette thèse concernant les comportements individuels de polarisation, une distinction entre métriques globales et individuelles est opérée.

1.4.1 Métriques globales de polarisation

En 1994, Esteban et Ray ont été les premiers à conceptualiser une mesure générale de polarisation [Esteban and Ray, 1994]. Pour cela, ils se basent essentiellement sur une division de la société en divers groupes, entraînant un accroissement de la polarisation. Plus précisément, à partir d'une distribution d'individus au sein de groupes $(\pi, y) = (\pi_1, ..., \pi_n, y_1, ..., y_n)$, où $\pi_i > 0$ correspond au poids du groupe i (*c.-à-d.* la taille du groupe) et y_i correspond à une valeur représentative (moyenne, médiane, *etc.*) d'un attribut dans le groupe i , par exemple l'opinion. La polarisation globale de cette distribution est évaluée selon trois fonctions : (1) une fonction d'identification I, traduisant le sentiment d'identification d'un individu vis-à-vis des individus proches de lui ; (2) une fonction d'aliénation a, traduisant le sentiment d'aliénation d'un individu vis-à-vis des individus éloignés de lui (distance δ) ; (3) une fonction d'antagonisme effectif $T(I, a)$, traduisant un lien entre le sentiment d'identification et celui d'aliénation. La polarisation totale P sur la population est alors définie comme la somme de tous ces antagonismes effectifs, selon l'Équation (1.4).

$$P(\pi, y) = \sum_i \sum_j \pi_i \pi_j T[I(\pi_i), a(\delta(y_i, y_j))] \tag{1.4}$$

où i et j sont des groupes. La polarisation est alors évaluée en fonction du degré d'homogénéité au sein des groupes, et d'hétérogénéité entre les groupes. Cette métrique fondamentale permet d'évaluer la polarisation au sens large, *c.-à-d.* en considérant la fragmentation d'individus au sein de groupes plus ou moins éloignés. Cependant, des métriques spécifiques à la polarisation en ligne, sur les médias et réseaux sociaux, ont été élaborées. Beaucoup d'entre elles sont calculées sur des graphes, permettant de représenter les échanges d'informations sur les médias sociaux [Beauguitte, 2010]. Un graphe $G = (N, L)$, est composé d'un ensemble de nœuds N et de liens L qui relient ces nœuds. Le graphe est défini par son ordre, *c.-à-d.* le nombre de nœuds, et sa taille, *c.-à-d.* le nombre de liens. Ces nœuds peuvent représenter une large variété d'informations en fonction du contexte d'application, comme des utilisateurs ou des sources d'information, et peuvent être de nature différente au sein d'un même graphe. De la même façon, les liens qui les relient peuvent correspondre à différentes formes d'interactions, comme la lecture, le suivi, le partage d'information ou encore des liens d'amitié [Nettleton, 2013]. De plus, ces liens peuvent être orientés et ainsi donner une information quant au sens de la relation entre les nœuds, mais aussi pondérés pour traduire l'intensité de la relation. Ces graphes peuvent être représentés visuellement, ou à l'aide d'une matrice d'adjacence. Quel que soit leur mode de représentation, les graphes sont des outils puissants et notamment largement utilisés pour l'analyse et la quantification de la polarisation [Adams *et al.*, 2022].

Dans le domaine des réseaux sociaux, [Conover *et al.*, 2011] représentent par exemple des réseaux de retweets et de mentions entre les utilisateurs à l'aide d'un graphe où chaque nœud représente un utilisateur, et chaque lien représente une interaction entre deux utilisateurs. Une

fois le graphe construit, les auteurs calculent la modularité [Newman, 2006], qui permet d'évaluer le partitionnement des nœuds d'un réseau (ou graphe) en différentes communautés en fonction de la présence de liens entre ces communautés. La modularité est définie dans l'Équation (1.5).

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \frac{s_i s_j + 1}{2} \tag{1.5}$$

où m correspond au nombre de liens, A est la matrice d'adjacence, k_i et k_j le degré des nœuds i et j (*c.-à-d.* le nombre de liens reliés au nœud) et $s_i s_j = 1$ si les nœuds i et j appartiennent à une même communauté, et $s_i s_j = -1$ sinon. Les valeurs de modularité ainsi calculées varient entre $-1/2$ et 1, et les valeurs supérieures à $0,3$ indiquent une partition marquée du réseau. Les valeurs de modularité permettent donc de refléter le partitionnement des utilisateurs en plusieurs communautés, au sein desquelles il y a une forte homogénéité, et entre lesquelles il y a une forte hétérogénéité. Cependant, [Guerra *et al.*, 2013] expliquent que la modularité ne semble pas être une mesure directe de la polarisation, car certains réseaux non polarisés peuvent également être divisés en communautés distinctes. Les auteurs soulignent également le rôle des nœuds placés aux limites des communautés (ou groupes) pour quantifier la polarisation. La présence de nœuds très fortement connectés aux intermédiaires des communautés identifiées traduit, selon eux, une polarisation plus faible. Ils proposent ainsi une métrique de polarisation comparant les connexions internes et externes de nœuds intermédiaires. Un nœud intermédiaire $v \in B$, où B est l'ensemble des nœuds intermédiaires, répond à deux conditions. Premièrement, il a un ensemble de connexions avec les nœuds internes à sa communauté, qui n'ont aucun lien avec une autre communauté. Le nombre de liens internes pour un nœud intermédiaire v est noté $d_i(v)$. Deuxièmement, il a au moins un lien avec un nœud d'une autre communauté. Le nombre de liens externes pour un nœud intermédiaire v est noté $d_b(v)$. La polarisation P est alors évaluée en fonction du nombre de liens internes et externes pour les nœuds intermédiaires dans un graphe, selon l'Équation (1.6).

$$P = \frac{1}{|B|} \sum_{v \in B} [\frac{d_i(v)}{d_b(v) + d_i(v)} - 0,5] \tag{1.6}$$

La valeur de polarisation résultante varie dans [-0,5, 0,5], avec $P < 0$ reflétant un manque de polarisation avec des nœuds intermédiaires plus susceptibles de faire des connexions avec une autre communauté, tandis que $P > 0$ indique qu'il y a de la polarisation et que les nœuds intermédiaires forment préferentiellement des liens au sein de leur communauté et $P = 0$.

Suivant une idée similaire, [Garimella *et al.*, 2018] proposent de quantifier la controversialité d'un sujet, *c.-à-d.* sa propension à provoquer des débats mouvementés. Les sujets très controversés conduisent automatiquement à des communautés bien séparées dans le réseau, et la controversialité peut donc être mesurée en examinant le flux d'informations entre chaque communauté. Les auteurs appliquent une marche aléatoire (*random walk*) dans le graphe, permettant d'évaluer la facilité pour un utilisateur à interagir avec un utilisateur d'une autre communauté.

Au-delà de l'utilisation de graphes, des mesures basées sur la distribution des opinions ou des avis ont été proposées dans la littérature. [Morales *et al.*, 2015] proposent par exemple de mesurer la polarisation en fonction de la distribution des opinions partagées dans un média social. Ainsi, plus les opinions sont inégalement distribuées et éloignées, plus la polarisation est importante. Dans un contexte où les utilisateurs ont l'opportunité d'attribuer des notes à des items ou des sujets, reflétant leur intérêt ou leur appréciation, [Badami *et al.*, 2017] proposent une mesure similaire de polarisation basée sur la distribution de ces notes : plus la distribution des notes est inégale, plus la polarisation est importante. Enfin, à partir de la distribution des

avis donnés par les utilisateurs, [Koudenburg *et al.*, 2021] proposent une mesure identique et la font valider en tenant compte de l'avis d'experts internationaux de la polarisation au travers d'une enquête.

Pour résumer, les mesures globales de polarisation présentent une caractéristique commune. Elles permettent de quantifier la polarisation de l'ensemble d'un réseau ou d'une thématique particulière sur la base d'un unique facteur de polarisation. Cependant, comme expliqué précédemment, la polarisation est influencée par de nombreux facteurs, dont certains peuvent être propres à chaque utilisateur, ce qui justifie l'existence de métriques individuelles.

1.4.2 Métriques individuelles de polarisation

Les métriques permettant une mesure individuelle de polarisation sont peu nombreuses. On trouve par exemple le *score de polarisation* proposé par [Becatti *et al.*, 2019], basé sur l'identification d'un ensemble de communautés C. Pour calculer ce score, un utilisateur u est représenté par un vecteur $I_{u,C}$ où chaque élément $I_{u,c}$ représente la proportion d'interactions de u avec la communauté c, $N_{u,c}$, par rapport à son nombre total d'interactions N_u. Le score de polarisation $\rho(u)$ de l'utilisateur u, est simplement évalué comme la valeur maximale du vecteur $I_{u,C}$, comme présenté dans l'Équation (1.7).

$$\rho(u) = \max_{c \in C} \{I_{u,c}\} \quad \text{with} \quad I_{u,c} = \frac{N_{u,c}}{N_u} \tag{1.7}$$

Les utilisateurs dont le *score de polarisation* est égal à 1 sont des utilisateurs qui accèdent à une communauté unique et sont donc fortement polarisés. Au contraire, les utilisateurs qui accèdent à toutes les communautés de manière égale ont un *score de polarisation* proche de $1/|C|$. La limite inférieure de ce score de polarisation dépend donc du nombre de communautés, ce qui limite sa qualité. En outre, puisqu'il est basé sur la valeur maximale du vecteur $I_{u,C}$, le *score de polarisation* ne prend en compte que la communauté avec laquelle l'utilisateur a le plus interagi. Ce score n'est pas orienté puisqu'il ne renseigne pas sur les communautés consultées, mais représente la mesure dans laquelle les interactions d'un utilisateur sont hétérogènes. Dans un contexte particulier opposant deux communautés principales, le score de polarisation présenté par [Schmidt *et al.*, 2018] est similaire mais varie entre -1 et 1 et est orienté : la valeur informe sur la communauté à laquelle l'utilisateur accède le plus souvent (Équation (1.8)).

$$\phi(u) = \frac{N_{u,c1} - N_{u,c2}}{N_{u,c1} + N_{u,c2}} \tag{1.8}$$

Avec $c1$ et $c2$ les deux communautés, $N_{u,c1}$ le nombre d'interactions de u dans $c1$, et $N_{u,c2}$ le nombre d'interactions de u dans $c2$. Un score proche de -1 indique que u est polarisé vers $c2$, tandis qu'un score proche de 1 indique que u est polarisé vers $c1$.

Plus récemment, [Cicchini *et al.*, 2022] ont proposé la métrique de *manque de diversité* (*Lack of Diversity* (LD)) qui est, dans une certaine mesure, très similaire au *score de polarisation* de Becatti *et al.* [Becatti *et al.*, 2019]. La principale différence réside dans le fait que cette métrique prend en compte les sources d'information avec lesquelles un utilisateur a interagi, concrètement un ensemble de M médias, plutôt que des communautés. Chaque utilisateur u est alors représenté par un vecteur $J_{u,M}$, où $J_{u,m}$ indique le nombre d'interactions $N_{u,m}$ de u sur les actualités des médias m. Cette composante est corrigée en fonction du nombre total d'interactions sur le média m. La formule de LD est donnée dans l'Équation (1.9).

$$LD(u) = \max_{m \in M} \{J_{u,m}\} \quad \text{with} \quad J_{u,m} = N_{u,m} \cdot log(\frac{|U|}{|U_m|}) \tag{1.9}$$

$|U|$ est le nombre total d'utilisateurs et $|U_m|$ est le nombre d'utilisateurs qui interagissent avec le média m. $log(\frac{U}{U_m})$ corrige un biais potentiel introduit par le média m lorsqu'il est partagé par un grand nombre d'utilisateurs. Comme dans [Becatti *et al.*, 2019], le score reflète la valeur maximale dans le vecteur propre à chaque utilisateur. Ainsi calculée, la métrique LD n'est pas bornée et doit donc être normalisée.

Bien que permettant une quantification de la polarisation pour chaque utilisateur, ces rares mesures individuelles de polarisation présentent certaines limites. Tout d'abord, elles ne sont calculées que sur un seul facteur (nombre d'interactions avec les communautés ou avec les médias). Par conséquent, en ne considérant qu'un unique facteur de polarisation, plusieurs utilisateurs peuvent être identifiés comme étant polarisés de la même manière selon ce dernier, mais peuvent en fait présenter un large éventail de comportements et se distinguer sur d'autres facteurs. Deux utilisateurs peuvent par exemple interagir avec une communauté commune et dans les mêmes proportions, mais consulter une diversité de sources bien différentes. Ensuite, le *score de polarisation* [Becatti *et al.*, 2019] et la mesure de *manque de diversité* [Cicchini *et al.*, 2022] n'exploitent que la valeur maximale du vecteur de l'utilisateur, les autres valeurs n'étant pas considérées dans le calcul final. Se baser uniquement sur la valeur maximale semble pourtant très réducteur : deux utilisateurs qui se comportent de la même façon dans leur communauté/média principal(e) obtiendront un même score, alors qu'ils peuvent se comporter de manière très différente dans d'autres communautés ou médias. Ces comportements hétérogènes ne sont pas pris en compte dans le calcul des métriques, ce qui limite la qualité de la modélisation. Dans ce travail, je ferai l'hypothèse que l'ensemble des interactions effectuées par un utilisateur apportent un éclairage utile, et contribuent à une meilleure compréhension du comportement de polarisation.

1.4.3 Métriques multi-factorielles de polarisation

Il n'existe que très peu de métriques capables de prendre en considération plusieurs facteurs lors de la mesure de la polarisation. Toutefois, [Phillips *et al.*, 2023] ont très récemment proposé une métrique multi-factorielle de polarisation. Elle repose sur trois dimensions : sociale, connaissance et sources. La dimension sociale représente la communication entre les membres d'une communauté. La connaissance correspond aux idées et arguments qui sont mis en avant dans la communauté. Enfin, les sources correspondent aux élites politiques ou organisations qui sont citées par la communauté. Ces dimensions sont évaluées sur la base d'un graphe G multi-dimensionnel : les nœuds représentés restent les mêmes pour évaluer chaque dimension, mais la nature et la position des liens varient en fonction du facteur de polarisation évalué. Les résultats montrent que la considération de divers facteurs lors de l'évaluation de la polarisation permet d'en tirer des conclusions plus complètes, et donc d'avoir une compréhension plus aboutie du processus de polarisation au sein d'un réseau. Cette métrique, bien que multi-factorielle, permet uniquement l'évaluation d'un ensemble et ne tient pas compte des différences inter-individuelles.

1.5 Approche temporelle et dynamique de la polarisation

À ma connaissance, la plupart des travaux de la littérature étudient la polarisation de façon statique, et peu d'entre eux étudient sa dynamique, correspondant à l'évolution temporelle des comportements de polarisation associés. Modélisée et évaluée statiquement, la polarisation est considérée comme un état de cristallisation des opinions, qui n'évolue pas [Norris, 2003]. Cependant, les résultats de la modélisation de cet état peuvent s'avérer complètement obsolètes dans le temps, car elle ne permet pas de considérer l'évolution temporelle des opinions. L'information partagée et discutée en ligne suit pourtant une certaine dynamique, avec une adaptation à

l'évolution des débats publics [Russell Neuman *et al.*, 2014].

En ce sens, quelques rares travaux se sont attelés à étudier l'évolution de la polarisation au cours du temps. [Garimella and Weber, 2017] ont par exemple montré une importante évolution de la polarisation sur Twitter à long terme, entre 2009 et 2016. Le niveau global de polarisation, évalué sur un panel de 679 000 utilisateurs est en effet passé de 10 à 20% en 7 ans, avec notamment des variations spécifiques au moment d'événements politiques aux États-Unis comme les élections présidentielles ou élections de mi-mandat. Cet effet du contexte politique sur l'évolution de la polarisation en ligne a également été mis en évidence dans le contexte suisse, où des variations du niveau de polarisation ont été identifiées à partir de données collectées sur la plateforme `Politnetz.ch`, qui permet la communication entre figures politiques du pays et électeurs [Garcia *et al.*, 2015]. Ainsi, le niveau global de polarisation semble évoluer temporellement, notamment sous l'effet d'événements contextuels, comme des élections politiques. Cette évolution semble cependant être différente en fonction du réseau social, où les contenus partagés et discutés ne sont pas les mêmes et engendrent des comportements différents [Yarchi *et al.*, 2021].

Ces résultats confirment l'intérêt de proposer une modélisation et une évaluation temporelle de la polarisation. [Pereira *et al.*, 2018] soulignent cependant sa complexité. Pour y répondre, ces auteurs proposent un modèle permettant d'évaluer l'évolution temporelle des préférences des utilisateurs sur les médias sociaux à l'aide de graphes temporels, plus adaptés à l'analyse des dynamiques de préférences. Au sein de ces graphes, des liens existants sont susceptibles de disparaître, et de nouveaux liens peuvent apparaître en fonction de l'évolution des interactions. Ainsi, les auteurs montrent une forte corrélation entre les modifications structurelles d'un graphe, avec notamment une évolution de la centralité des nœuds, et l'évolution des préférences des utilisateurs. Dans une analyse plus récente, [Tardelli *et al.*, 2023] s'intéressent aux changements temporels des communautés identifiées sur les médias sociaux, et mettent en évidence l'existence de certains groupes d'utilisateurs coordonnés qui suivent des trajectoires similaires au cours du temps.

L'ensemble de ces résultats souligne l'intérêt de s'intéresser à l'évolution temporelle de la polarisation, permettant d'en avoir une compréhension plus approfondie que celle offerte par une modélisation statique de la polarisation. Cependant, à notre connaissance, ces travaux adoptent à nouveau une approche globale, et étudient la polarisation sur l'ensemble des utilisateurs. Selon moi, aborder l'évolution temporelle des comportements de polarisation individuels permettrait pourtant d'avoir une modélisation plus précise, nécessaire au développement d'approches de dépolarisation personnalisées.

1.6 Synthèse et positionnement

Dans le contexte informationnel actuel, de nombreux travaux étudient, modélisent et mesurent la polarisation. Ces derniers montrent que le format des réseaux sociaux contribue à exacerber la polarisation, et à la co-construire artificiellement, notamment sous l'effet d'une exposition sélective accrue.

L'objectif de notre travail n'est pas de remettre en cause l'impact des médias sociaux sur la polarisation [Kubin and Von Sikorski, 2021, Prior, 2013], en particulier en ce qui concerne le filtrage algorithmique de l'information qui tend à conforter les utilisateurs dans leurs croyances [Pariser, 2011]. Néanmoins, les principaux écueils des travaux de la littérature sont :

— La modélisation uni-dimensionnelle communément appliquée ne permet pas de tenir compte de la multiplicité des facteurs influençant l'apparition de la polarisation.

— Les mesures individuelles de polarisation existantes ne tiennent compte que des comporte-

ments majoritaires et ne permettent pas d'évaluer pleinement la polarisation.

— La modélisation des dynamiques temporelles de polarisation est globale, et ne permet pas d'étudier l'évolution des comportements individuels sous-jacents.

Idéalement, les approches de la littérature devraient être adaptables à tous les contextes d'étude et à toutes les données utilisées. Ainsi, dans une perspective de réduction et de contrôle du phénomène de polarisation pour en éviter les dérives, il me semble nécessaire de pouvoir quantifier la polarisation de façon individuelle et multi-factorielle. Cela s'accompagne d'une modélisation plus fine du comportement de polarisation afin d'en identifier les différents aspects et d'évaluer son évolution temporelle (Figure 1.2).

Les objectifs de travail permettant de répondre aux limites identifiées de la littérature, ainsi qu'à ma première question de recherche (*QR1*), sont donc les suivants :

— Proposer une modélisation individuelle et multi-factorielle du phénomène de polarisation

— Proposer une modélisation temporelle du phénomène de polarisation

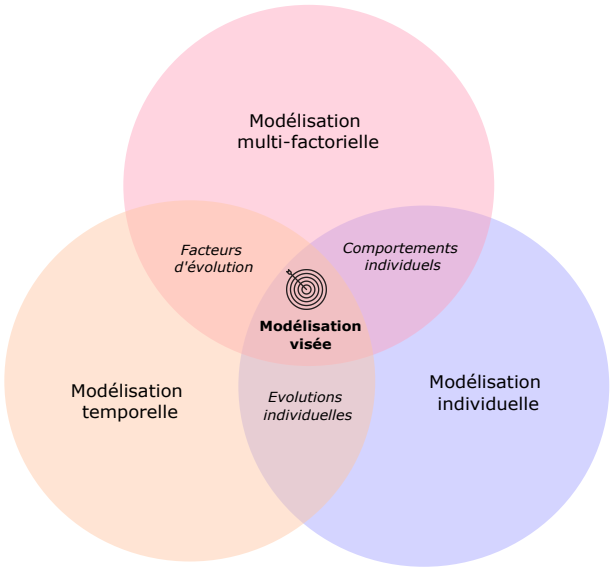


FIGURE 1.2 – Positionnement du travail - Modélisation

Chapitre 2

Modélisation individuelle et multi-factorielle du phénomène de polarisation

Sommaire

2.1 Introduction

Ce premier chapitre de contribution vise à répondre au manque de modélisation fine du phénomène de polarisation dans la littérature, résultant principalement de l'application de modèles globaux sur un ensemble d'utilisateurs. La quantification de la polarisation résultant de cette modélisation est non représentative des comportements de polarisation adoptés car elle se positionne à un niveau global. Ainsi, je propose de travailler sur la définition d'une métrique individuelle et multi-factorielle de polarisation, dans le but de contribuer à une modélisation plus fine

du phénomène. Les travaux présentés répondent ainsi à l'objectif de proposer une modélisation individuelle et multi-factorielle du phénomène de polarisation. Je pose donc l'hypothèse selon laquelle cette modélisation offre une quantification pertinente du phénomène de polarisation, essentielle pour élaborer des solutions adaptées visant à le réduire. Plus précisément, une modélisation plus fine permettrait de discriminer divers comportements de polarisation, nécessitant des réponses spécifiques pour diminuer la polarisation.

Les sous-questions de recherche découlant de cet objectif sont :

> ***Sous-questions de recherche abordées dans ce chapitre :***
> Comment modéliser et combiner de multiples facteurs dans une métrique individuelle de polarisation ? (*QR1.1*)
>
> Une telle combinaison permet-elle d'expliquer les comportements de polarisation adoptés ? (*QR1.2*)

Dans la suite de ce chapitre, la métrique proposée, nommée GRAIL (GeneRalized AddItive poLarization), est introduite. Dans un second temps, le protocole expérimental permettant d'évaluer cette métrique est détaillé. Finalement, les résultats de cette validation expérimentale sont présentés.

2.2 GRAIL : une métrique individuelle et multi-factorielle de polarisation

L'objectif de cette section est de proposer une nouvelle métrique de polarisation répondant aux limites des métriques existantes de la littérature. Ainsi, la métrique proposée doit permettre d'**évaluer finement la polarisation au travers d'une mesure individuelle et multi-factorielle pour rendre compte de la complexité du phénomène de polarisation.**

Les différentes étapes ayant conduit à la définition d'une telle métrique sont détaillées dans la suite de cette section. Pour guider la lecture de cette dernière et illustrer les éléments qui constituent la métrique proposée, les explications sont ponctuées d'un exemple fil rouge.

2.2.1 Modéliser finement un facteur de polarisation

La principale limite des métriques individuelles de polarisation de la littérature réside dans la quantification partielle de la polarisation pour chaque utilisateur. En effet, quel que soit le facteur de polarisation évalué, la communauté d'intérêt avec le score de polarisation ρ (Équation 1.7), ou les sources d'information avec la métrique de manque de diversité LD (Equation 1.9), seules les interactions majoritaires d'un utilisateur sont exploitées pour évaluer la polarisation. Selon moi, cela **limite la qualité de la quantification de la polarisation car la mesure ne prend pas en compte l'ensemble des interactions, qui apportent pourtant une information importante sur le comportement adopté par l'utilisateur**.

Pour répondre à cette limite, je propose d'évaluer un facteur de polarisation à l'aide d'une mesure issue du domaine de la théorie de l'information, l'entropie [Shannon, 1948]. Etant donné que la valeur maximale de l'entropie, telle qu'initialement définie, n'est pas bornée, je propose d'utiliser l'entropie normalisée afin d'obtenir une valeur appartenant à l'intervalle $[0; 1]$. Par souci de simplicité, cette entropie normalisée est appelée entropie dans la suite de ce manuscrit.

L'entropie

L'entropie est une mesure de l'incertitude, ou de la quantité d'information dans un ensemble de données, dont la formule est détaillée dans l'Équation (2.1).

$$H_N(Z) = \frac{-\sum_z P(z)log(P(z))}{log(s)} \tag{2.1}$$

où Z est une variable aléatoire discrète qui prend s valeurs d'entités possibles, et $P(z)$ est la probabilité de l'entité z. L'entropie peut être appliquée à tout facteur pouvant être considéré comme une variable aléatoire.

La valeur d'entropie est maximale lorsqu'il y a équiprobabilité, *c.-à-d.* que la masse des probabilités est distribuée de façon homogène. Inversement, si une partie importante de la masse des probabilités est répartie sur un petit ensemble d'entités, l'entropie est faible.

L'entropie pour modéliser un facteur de polarisation

Dans mon contexte de recherche sur la polarisation, je propose d'appliquer cette mesure d'entropie pour quantifier la dispersion des interactions d'un utilisateur. **La valeur d'entropie calculée informe ainsi sur la diversité des interactions d'un utilisateur**. Ainsi, soit Z un facteur composé d'un ensemble d'entités avec lesquelles les utilisateurs peuvent interagir. Ce facteur peut être représenté sous la forme d'une distribution de probabilité afin de représenter la dispersion des interactions d'un utilisateur avec ces entités. Le calcul d'entropie appliqué à cette distribution permet ainsi de quantifier la dispersion des interactions de l'utilisateur sur ces entités : plus l'entropie est élevée, plus l'utilisateur interagit avec des entités diversifiées. Dans le contexte d'étude de la polarisation, les variables aléatoires peuvent par exemple être les sujets d'actualité abordés dans des news, les sources d'information ou médias, les comptes suivis, les opinions.

Appliquées à cette même distribution de probabilité, les métriques individuelles de polarisation de la littérature, ρ et LD (Équation (1.7), page 24 et Équation (1.9), page 24), ne considèrent que la probabilité maximale et ignorent les autres probabilités. Par conséquent, deux utilisateurs différents ayant la même probabilité maximale, mais des distributions de probabilités différentes pour les autres entités, obtiendraient des valeurs de polarisation identiques. En revanche, en tenant compte de l'ensemble des interactions, le calcul de l'entropie produirait des valeurs distinctes. Ainsi, je fais l'hypothèse que cette approche basée sur l'entropie offre une modélisation plus fine des facteurs de polarisation par rapport à celle proposée dans la littérature.

Pour confirmer le potentiel de l'entropie pour l'évaluation de la polarisation, j'introduis maintenant l'exemple fil rouge qui sera utilisé tout au long de cette section. Z est l'ensemble de sources d'information avec lesquelles les utilisateurs peuvent interagir et $s = 4$ sources. En fonction des interactions des utilisateurs avec chacune de ces sources, une distribution de probabilité peut être calculée pour chaque utilisateur u et forme un vecteur de probabilité de taille s, avec $P_u(z)$ la probabilité que l'utilisateur u interagisse avec la source z. Le Tableau 2.1 présente un exemple de distributions de probabilités d'interactions avec les sources Z pour quatre utilisateurs u_1 à u_4. Les scores résultants calculés avec les métriques individuelles de polarisation de la littérature, ρ et LD, ainsi que l'entropie normalisée, définie dans l'Équation (2.1), peuvent donc être calculés pour chaque utilisateur et sont également indiqués dans le Tableau 2.1. Pour simplifier cet exemple, il est supposé que les quatre sources z ont toutes

la même popularité. Le terme correcteur de LD (Équation (1.9), page 24) est donc égal à 1, et $\rho = LD$.

Utilisateur	**Distribution de probabilité**	ρ **et** LD	$H_N(Z)$	$f(H_N(Z))$
u_1	0.5 | 0 | 0.5 | 0	0.50	0.50	0.50
u_2	0.5 | 0.2 | 0.2 | 0.1	0.50	0.88	0.73
u_3	0.3 | 0.3 | 0.3 | 0.1	0.30	0.95	0.81
u_4	0.25 | 0.25 | 0.2 | 0.3	0.30	0.99	0.90

TABLE 2.1 – Exemple de distribution de probabilité pour 4 utilisateurs adoptant des comportements différents

Comme attendu, pour les utilisateurs u_1 et u_2, dont la probabilité maximale est identique $(0,5)$, le niveau de polarisation est le même selon les métriques de la littérature, avec $\rho = LD = 0,5$. Ces deux utilisateurs sont donc associés à un même niveau de polarisation, alors qu'ils interagissent très différemment sur les sources d'information disponibles. Au contraire, en quantifiant la diversité de leurs interactions avec l'entropie, qui tient compte de l'ensemble de la distribution et considère chaque entité du vecteur, les utilisateurs u_1 et u_2 ont des valeurs bien différentes avec $H_N(Z) = 0,50$ pour u_1 et $H_N(Z) = 0,88$ pour u_2. La valeur d'entropie plus élevée pour l'utilisateur u_2 informe d'une diversité plus importante dans les sources d'information auxquelles il a accédé. Cela est vérifié puisqu'u_2 interagit avec les quatre sources d'informations disponibles, tandis qu'u_1 n'interagit qu'avec deux d'entre elles. Cet exemple illustre que **l'utilisation de l'entropie, en tenant compte de l'ensemble des dimensions du vecteur de probabilité, permet de caractériser différemment les interactions par rapport aux métriques de la littérature**.

2.2.2 Différencier les comportements de polarisation

L'une des caractéristiques de la métrique que je souhaite concevoir est de permettre une meilleure discrimination entre les utilisateurs. L'utilisation de l'entropie, présentée dans la sous-section précédente, est un premier pas dans cette direction. Bien qu'allant au-delà des métriques individuelles de polarisation de la littérature en considérant l'ensemble des interactions des utilisateurs, les valeurs d'entropie calculées peuvent être proches pour des utilisateurs adoptant des comportements différents, ce qui limite leur différenciation. Dans l'exemple fil rouge introduit, les utilisateurs u_3 et u_4 ont par exemple des valeurs d'entropie proches $(0,95$ et $0,99$ respectivement), mais adoptent des comportements différents puisque les interactions de u_4 sont réparties de façon plus homogène sur les sources d'informations Z. L'objectif est donc de **discriminer plus finement les utilisateurs de façon à différencier les comportements de polarisation.**

Pour répondre à cela, je propose d'appliquer une transformation des valeurs d'entropie de l'ensemble des utilisateurs afin de modifier leur distribution dans l'espace de représentation, et contribuer à mieux les différencier. Plus particulièrement, je propose d'appliquer une transformation non linéaire permettant idéalement de discriminer les utilisateurs pour qui les valeurs d'entropie sont très proches.

La fonction sigmoïde

Une fonction répondant à ce besoin est la fonction logistique sigmoïde, modélisant une relation non linéaire entre des variables (Équation (2.2)).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.2}$$

Cette fonction logistique est bornée, augmente de façon monotone, et sa courbe caractéristique en forme de S possède exactement un point d'inflexion au centre et à équidistance des deux extrémités (Figure 2.1). Intuitivement, ce type de courbe répond au besoin de différentiation visé avec une phase exponentielle utile pour différencier les utilisateurs aux valeurs d'entropie trop proches. Par ailleurs, une phase transitoire permet de conserver la distribution naturelle des données pour certaines valeurs, et une phase stationnaire permettant de rassembler les utilisateurs ayant un comportement similaire, sans introduire de biais si ce comportement est trop fréquent.

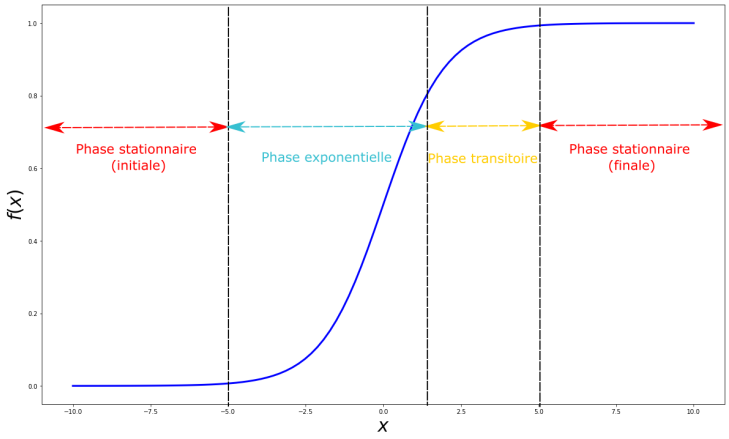


FIGURE 2.1 – Fonction sigmoïde

Une fonction polynomiale adaptée pour différencier les utilisateurs

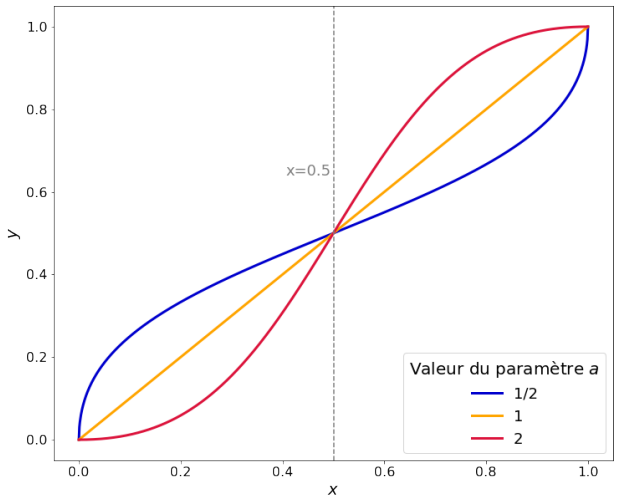
Bien que la fonction sigmoïde semble permettre de répondre à l'objectif de différenciation des utilisateurs pour qui les valeurs sont très proches, elle présente une limite majeure en possédant deux asymptotes avec les équations $y = 0$ et $y = 1$. Dans le cas de la quantification de la polarisation, la métrique que je construis doit être définie de façon à ce que les valeurs varient dans $[0; 1]$, 0 indiquant une absence de polarisation et 1 une polarisation extrême, conformément aux mesures de la littérature [Becatti *et al.*, 2019, Cicchini *et al.*, 2022]. L'objectif est donc d'identifier une fonction f numériquement stable avec $f(0) = 0$ et $f(1) = 1$, et ayant un profil similaire à celui de la fonction sigmoïde. Inspirée de l'équation de la fonction sigmoïde (Équation (2.2)), la fonction polynomiale définie dans l'Équation (2.3) permet de répondre à ces contraintes.

$$f(x) = \frac{x^a}{x^a + (1 - x)^a} \tag{2.3}$$

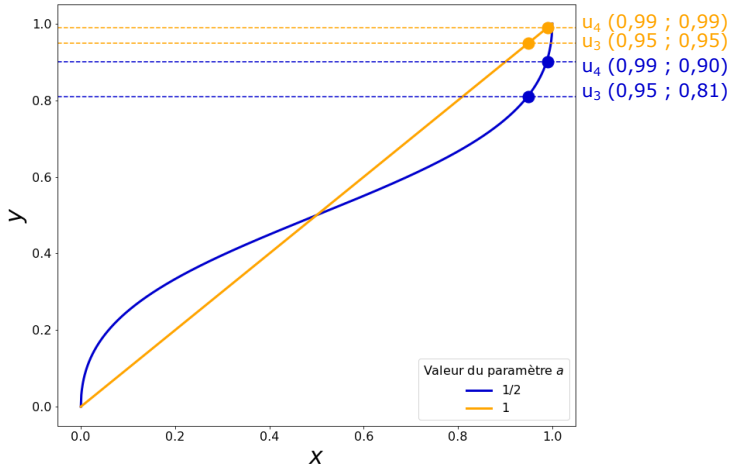
Dans cette formule, x^a remplace le terme exponentiel e^{-x} de la fonction sigmoïde, et contrôle la forme de la courbe. Plus précisément, le paramètre a influence à la fois la forme et la sensibilité de la pente de la courbe. En effet, lorsque $a > 1$, plus a est élevé, plus la pente de la courbe est raide et plus la fonction est sensible aux petites variations des valeurs intermédiaires de x.

Cependant, lorsque $a < 1$, la courbe devient plus sensible aux variations des valeurs extrêmes de x (proches de 0 et de 1). Finalement, si $a = 1$, la fonction est linéaire (voir Figure 2.2a).

Je reviens ici à l'exemple fil rouge. L'équation (2.3) est appliquée à la valeur d'entropie de chaque utilisateur. Tout d'abord, la transformation polynomiale n'a aucun impact sur u_1, quelle que soit la valeur du paramètre a, puisque l'entropie de u_1 est de $0,5$. En outre, en fixant la valeur du paramètre a à $a = 1/2$, la valeur de l'entropie transformée de u_2, u_3 et u_4 est indiquée dans la dernière colonne du Tableau 2.1. Il convient de noter que u_3 et u_4, **les deux utilisateurs ayant des valeurs d'entropie extrêmes et similaires, sont mieux distingués une fois la fonction polynomiale proposée appliquée** (Figure 2.2b). Les utilisateurs u_1 et u_2, pour qui les valeurs d'entropie était bien différentes ($0,50$ et $0,88$), sont rapprochés mais restent suffisamment éloignés pour pouvoir être distingués.



(a) Variations en fonction du paramètre a.



(b) Application aux utilisateurs exemples u_3 et u_4.

FIGURE 2.2 – Fonction polynomiale.

En maîtrisant à la fois la forme et la sensibilité de la fonction polynomiale appliquée grâce au paramètre a, il est possible d'adapter la transformation en fonction de la nature des données étudiées ainsi que la différenciation des utilisateurs.

2.2.3 Combiner les indicateurs de polarisation dans une métrique unique

Comme expliqué dans l'état de l'art la polarisation peut se produire dans de multiples contextes et être influencée par de nombreux facteurs. L'objectif, annoncé dès le début de ce manuscrit, est d'être en mesure de **tenir compte de ces multiples facteurs dans la quantification de la polarisation, afin d'en avoir une meilleure compréhension.**

Ces facteurs ne sont pas nécessairement liés de manière linéaire, car ils peuvent observer des variations instables, et leur relation peut être délicate à modéliser. Par exemple, la diversité des sources d'information auxquelles un utilisateur spécifique accède peut ne pas être en relation linéaire avec la diversité des sujets qui l'intéressent. À cet égard, les modèles additifs généralisés, ou *Generalized Additive Model* (GAM), sont des modèles statistiques appropriés pour combiner plusieurs facteurs dans le calcul d'une métrique unique [Hastie, 2017] (Équation (2.4)).

$$GAM = \sum_i \alpha_i f_i(X_i) + \beta \text{ avec } \quad \sum_i \alpha_i = 1 \tag{2.4}$$

où f_i est une fonction régulière, X_i est une variable prédictive, α_i est le poids de la variable prédictive X_i et β est le terme représentant un potentiel biais. Les GAM permettent ainsi de modéliser des relations complexes entre une variable réponse Y et plusieurs variables prédictives X_i. La relation entre chaque variable X_i et Y est modélisée à l'aide d'une fonction complexe appelée *fonction régulière* (*smooth function*) f_i. Ces fonctions régulières sont souvent représentées par des *splines*, qui sont des fonctions définies par morceaux par des polynômes. Elles peuvent être combinées dans un modèle unique ou être ajoutées une par une pour chaque variable prédictive. Dans l'objectif de modélisation du phénomène de polarisation, la relation modélisée par le GAM est donc la relation entre les facteurs de polarisation X_i, liés au comportement des utilisateurs, et la polarisation Y.

2.2.4 GRAIL : une métrique individuelle et multi-factorielle de polarisation

Définition de GRAIL

Les différents composants décrits dans les sous-sections précédentes, à savoir (1) l'**entropie** permettant de modéliser un facteur de polarisation, (2) la **transformation polynomiale** pour différencier les utilisateurs, et (3) le **modèle additif généralisé** pour combiner plusieurs facteurs permettent de **construire une métrique individuelle et multi-factorielle de polarisation, la métrique GRAIL**.

Pour l'application de cette métrique, un facteur de polarisation X_i est donc évalué pour chaque utilisateur individuel u au travers de l'entropie, et plus précisément par $H'(X_{u,i}) = 1 - H_N(X_{u,i})$ pour que des valeurs plus élevées reflètent une plus forte polarisation. Ces facteurs X_i sont ensuite transformés de façon à différentier les utilisateurs à l'aide de la fonction polynomiale définie dans l'Équation (2.3). La fonction résultante est présentée dans l'Équation (2.5).

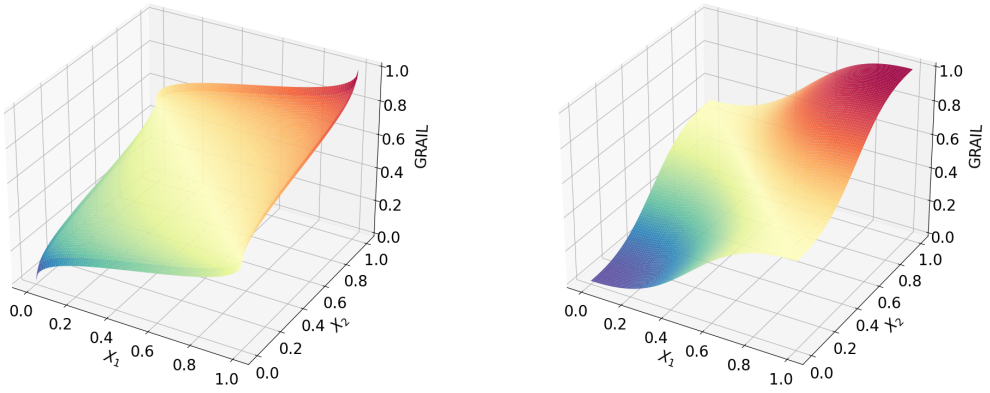
$$f_i(X_{u,i}) = \frac{H'(X_{u,i})^{a_i}}{H'(X_{u,i})^{a_i} + (1 - H'(X_{u,i}))^{a_i}} \tag{2.5}$$

La relation entre chacun des facteurs de polarisation $X_{u,i}$ et la variable réponse Y_u (*c.-à-d.* la polarisation de l'utilisateur u), est modélisée par un GAM, avec le terme biais $\beta = 0$ pour que les valeurs restent comprises dans $[0; 1]$. Par ailleurs, si $\alpha_i = \alpha_j \ \forall (i, j)$, cela revient à un GAM traditionnel, sans pondération. La formule finale de la métrique GRAIL est donnée dans l'Équation (2.6).

$$GRAIL\ (u) = \sum_i \alpha_i f_i(X_{u,i}) \tag{2.6}$$

$$\text{avec} \quad \sum_i \alpha_i = 1$$

Les variations des valeurs de GRAIL lors de la considération de 2 facteurs de polarisation, et selon différentes valeurs du paramètre a de la fonction polynomiale, sont présentées dans la Figure 2.3.



(a) $a = 0,5$ (b) $a = 2$

FIGURE 2.3 – Variations des valeurs de GRAIL lorsqu'appliquée à 2 facteurs et avec différentes valeurs du paramètre a.

Orientation de la métrique GRAIL

Suivant l'intuition de la métrique de [Schmidt *et al.*, 2018], présentée dans l'état de l'art (Équation (1.8), page 24) qui est adaptée à un contexte de polarisation affrontant deux communautés, je propose d'orienter la métrique GRAIL. En effet, les sujets fortement polarisés confrontent souvent deux communautés opposées : pro-vie/anti-vie, pro-armes/anti-armes, pro-vaccins/anti-vaccins, gauche/droite, *etc.* En l'orientant, la métrique GRAIL peut ainsi informer non seulement sur le degré de polarisation d'un utilisateur, mais aussi sur sa communauté d'appartenance et donc sur la nature de sa polarisation.

La transformation proposée pour orienter GRAIL est présentée dans l'Équation (2.7) :

$$GRAIL^{\pm}(u) = sgn(p) * \sum_i \alpha_i f_i(X_{u,i}) \tag{2.7}$$

$$\text{avec} \quad \sum_i \alpha_i = 1,$$

$$sgn(p) = \begin{cases} -1 & \text{si } u \text{ appartient à une communauté donnée} \\ +1 & \text{si } u \text{ appartient à la communauté opposée} \end{cases}$$

Pour revenir à l'exemple fil rouge, des facteurs de polarisation complémentaires à la diversité des sources pourraient être modélisés pour les utilisateurs u_1 à u_4. Par exemple, des indicateurs tels que la diversité des sujets ou des communautés avec lesquels les utilisateurs ont interagi compléteraient la modélisation. Ainsi, notons X_1 le facteur relatif aux sources d'information présenté dans l'exemple fil rouge depuis le début de cette section, et X_2 le facteur relatif aux communautés avec lesquelles les utilisateurs peuvent interagir, avec $s = 2$ communautés opposées. L'entropie peut alors être calculée pour ce second facteur, puis ces valeurs sont transformées selon la transformation polynomiale proposée. Fixons les paramètres de GRAIL à $a = 2$ et $\alpha = 0,4$ pour cet exemple, la valeur de polarisation GRAIL orientée pour chaque utilisateur u est alors calculée selon la formule suivante (Équation (2.8)) :

$$GRAIL^{\pm}(u) = sgn(p) \times \left(0,4 \frac{H'(X_{u,1})^2}{H'(X_{u,1})^2 + (1 - H'(X_{u,1}))^2} + 0,4 \frac{H'(X_{u,2})^2}{H'(X_{u,2})^2 + (1 - H'(X_{u,2}))^2} \right) \tag{2.8}$$

Je souhaite ajouter ici que les variables prédictives X_i, ainsi que la fonction régulière f_i, peuvent être instanciées par toute autre équation prenant ses valeurs dans $[0; 1]$ pour s'adapter à tout contexte de recherche et aux données associées. De plus, les paramètres de GRAIL peuvent être optimisés pour s'adapter au mieux aux données étudiées. Premièrement, le paramètre α peut être fixé pour correspondre à certains besoins spécifiques d'un contexte de recherche particulier, ou pour refléter des comportements spécifiques observés dans un ensemble de données. Par exemple, si un facteur de polarisation est majoritaire, ce dernier peut être sur-pondéré en adaptant la valeur du paramètre α. Deuxièmement, le paramètre a de la fonction polynomiale peut également être ajusté en fonction de la distribution des valeurs des données. Par exemple, si de nombreux utilisateurs ont des valeurs intermédiaires, $a > 1$ permettra de mieux les discriminer, tout en rassemblant les utilisateurs ayant des valeurs extrêmes. Au contraire, dans un contexte avec peu de valeurs intermédiaires, $a < 1$ devrait être préféré afin de garder un pouvoir de différenciation sur les valeurs extrêmes.

Je souhaite préciser ici que la phase d'optimisation du paramètre a peut également servir à évaluer le pouvoir polarisant d'un sujet au sein de l'ensemble de données étudié. La phase d'optimisation de ces paramètres peut être basée sur des observations ou bien sur des connaissances expertes dans un environnement d'étude contrôlé. De manière optimale, plusieurs fonctions objectif peuvent être définies et les valeurs des paramètres fixées en fonction de leur optimisation.

Pour résumer, l'ensemble des constituants formant la métrique GRAIL permettent d'en faire une métrique :

- **(C1) Individuelle** puisqu'elle permet de quantifier la polarisation à partir de l'ensemble des interactions de chaque utilisateur ;
- **(C2) Multi-factorielle** car elle permet de combiner divers facteurs ;
- **(C3) Discriminante** car elle permet d'aller au-delà de la simple distinction polarisé/non polarisé proposée par les travaux de la littérature ;
- **(C4) Généralisable** car elle peut être optimisée et adaptée en fonction des données exploitées.

Cette métrique GRAIL informe ainsi quant au **degré et à la nature de la polarisation de chaque utilisateur, en fonction de multiples facteurs**. Le degré correspond à l'intensité de la polarisation, tandis que la nature renseigne sur la communauté dont un utilisateur est le

plus proche ou à laquelle il appartient. Selon moi, l'application de cette métrique participe à une modélisation plus fine du phénomène de polarisation.

La section suivante présente le protocole expérimental permettant d'évaluer dans quelle mesure la quantification de la polarisation permise par GRAIL participe à une modélisation fine du phénomène de polarisation.

2.3 Protocole expérimental

2.3.1 Objectif

L'objectif de la validation expérimentale présentée dans ce chapitre est tout d'abord d'évaluer les différents composants de la métrique GRAIL, composant par composant, puis d'évaluer la pertinence de la quantification de la polarisation permise par GRAIL. La proposition de cette métrique individuelle et multi-factorielle de polarisation est guidée par un besoin de modélisation plus fine de la polarisation, et en particulier des comportements de polarisation adoptés par les utilisateurs. Ainsi, la sous-question de recherche à laquelle je cherche à répondre au travers de cette évaluation est (*QR1.2*) : La combinaison proposée par GRAIL permet-elle d'expliquer les comportements de polarisation adoptés ?

2.3.2 Description du protocole

Pour répondre à cette question, je propose donc de comparer les comportements de polarisation qui peuvent être différenciés et caractérisés à partir de la métrique GRAIL, à ceux pouvant être différenciés et caractérisés à partir des métriques individuelles de la littérature. Ces métriques baseline sont **le score de polarisation** ρ décrit par [Becatti *et al.*, 2019] (Équation (1.7), page 24) et **le manque de diversité** LD décrit par [Cicchini *et al.*, 2022] (Équation (1.9), page 24). Je rappelle que ces deux métriques sont construites de façon similaire, en ne considérant que les interactions maximales de chaque utilisateur.

Dans un premier temps, les facteurs de polarisation modélisés au travers de la métrique GRAIL, reposant sur un calcul d'entropie (Section 2.2.1) puis une transformation polynomiale (Section 2.2.2), sont évalués individuellement. Cette modélisation de chaque facteur est comparée à celle permise par l'application des métriques baseline. Une fois que la modélisation des facteurs au travers des composants de GRAIL est évaluée, c'est la pertinence de la combinaison de ces facteurs au sein d'une métrique unique pour la quantification de la polarisation que je cherche à valider. Les différentes étapes de ce protocole d'évaluation sont résumées dans la Figure 2.4.

Étape 1 : modélisation des facteurs de polarisation

La première étape du protocole d'évaluation consiste à comparer les distributions des facteurs parmi l'ensemble des utilisateurs lorsqu'ils sont modélisés avec l'entropie ou avec les métriques baseline. L'étude de cette distribution permet d'évaluer comment les valeurs sont distribuées parmi les utilisateurs, si certaines valeurs sont sur-représentées, ou au contraire sous-representées, en fonction de la façon dont sont évalués les facteurs.

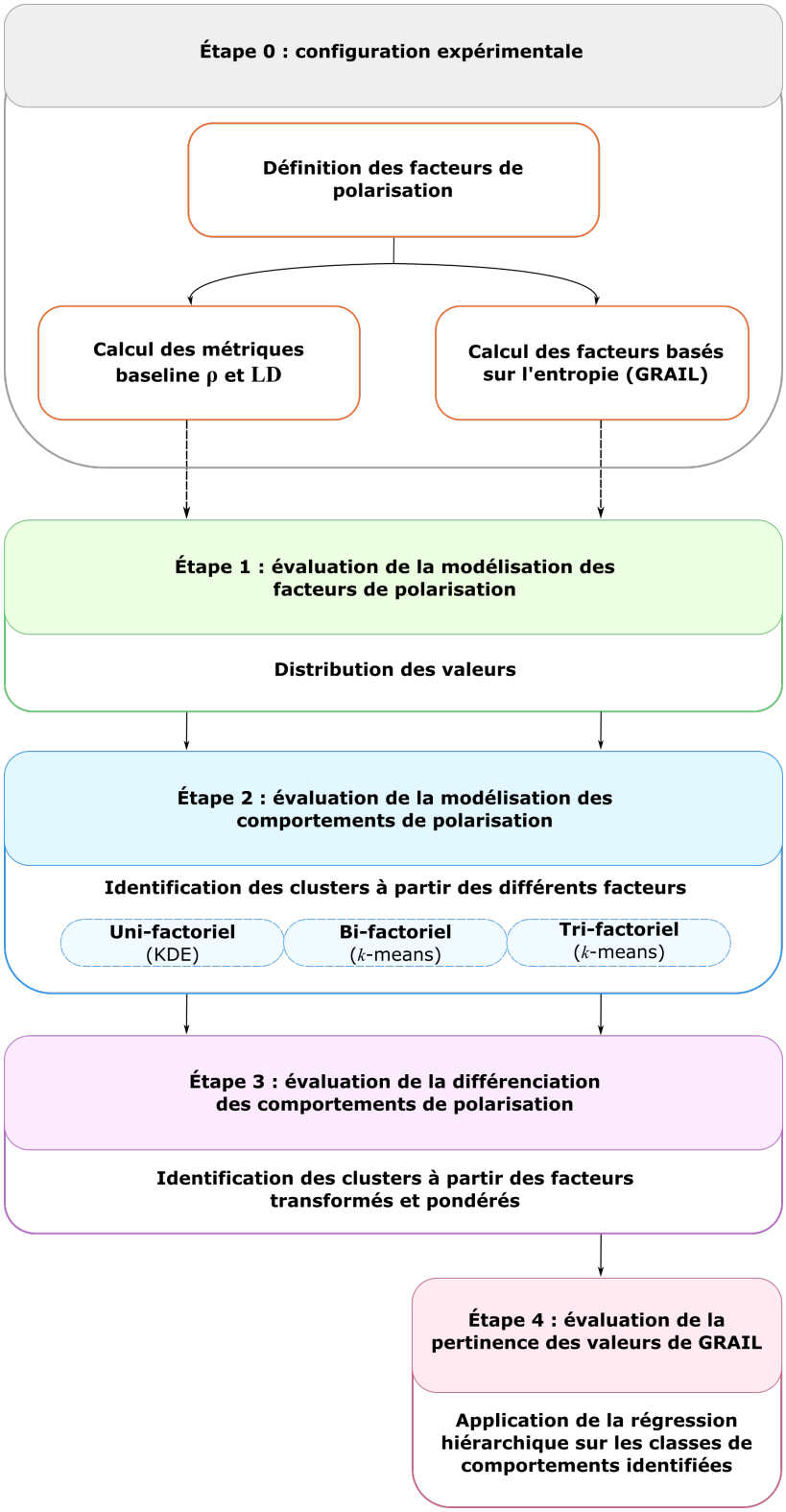


FIGURE 2.4 – Protocole d'évaluation de la métrique GRAIL

Étape 2 : modélisation des comportements de polarisation

La seconde étape du protocole vise à la fois à comparer la qualité de la modélisation des comportements de polarisation avec les baselines ou par les facteurs de GRAIL, et à évaluer l'apport d'une approche multi-factorielle.

Dans un premier temps, lorsque les facteurs sont considérés indépendamment, je propose de modéliser les comportements de polarisation à l'aide de l'approche de l'estimation de la densité du noyau, KDE (*Kernel Density Estimation*). Cette approche permet d'estimer la distribution de probabilité continue d'un facteur de polarisation dans un espace unidimensionnel. Appliquée aux facteurs de polarisation, cette distribution donne des indications quant à la répartition des valeurs du facteur unique étudié. Cette distribution permet ensuite d'identifier les minima et maxima locaux, à partir desquels peuvent être définis des ensembles de données distincts.

Sur l'exemple de la Figure 2.5, les minima sont en rouge et les maxima en vert, et permettent d'identifier 3 ensembles de valeurs représentés par les segments de couleur rouge, vert et bleu. Selon moi, dans le contexte d'étude sur la polarisation, les ensembles de données ainsi identifiés peuvent être interprétés comme des ensembles homogènes d'utilisateurs adoptant un comportement similaire en fonction du facteur de polarisation évalué.

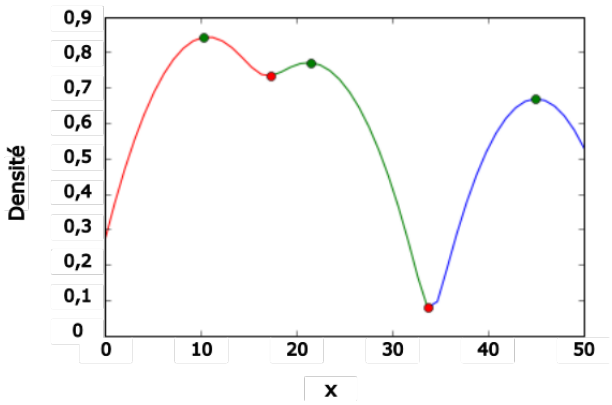


FIGURE 2.5 – Exemple de courbe de densité et des minima (en rouge) et maxima (en vert) locaux identifiés.

Dans un second temps, je propose d'évaluer plusieurs facteurs simultanément, et d'évaluer la modélisation des comportements de polarisation au travers d'une tâche de clustering, dont l'objectif est d'identifier des clusters d'utilisateurs à partir des facteurs de polarisation évalués. Ces clusters représentent ainsi des ensembles homogènes d'utilisateurs en fonction des indicateurs de polarisation. **L'application d'un algorithme de clustering permet ainsi d'identifier différents clusters, que je caractérise de classes de comportement de polarisation.** Pour l'identification de ces clusters, je propose d'utiliser l'algorithme de clustering k-means [Likas *et al.*, 2003]. Je sélectionne cet algorithme non supervisé car il est adapté aux données numériques et à faible dimension, mais aussi pour son faible coût de calcul. Le nombre optimal de clusters k, qui doit être fixé *a priori*, est optimisé en maximisant deux métriques de performance de clustering :

— **L'indice de Silhouette** [Rousseeuw, 1987] qui calcule, pour chaque donnée, la différence entre la distance moyenne avec les points du même cluster et la distance moyenne avec les points des clusters voisins. L'indice de Silhouette final est la moyenne des valeurs pour tous les points du jeu de données qui ont été classifiés. Les valeurs varient dans $[-1; 1]$, avec des valeurs négatives indiquant que les points sont en moyenne plus proches des points

de clusters voisins, et des valeurs positives indiquant que les points sont en moyenne plus proches des points de leur propre cluster. Un indice de Silhouette égal à 1 indique donc la classification optimale.

— **L'indice de Davies-Bouldin** [Davies and Bouldin, 1979] qui repose sur le principe qu'une classification correcte est associée à une faible variation au sein d'un cluster et une séparation élevée entre les clusters. L'indice calculé correspond ainsi à la moyenne du rapport maximal entre la distance d'un point au centre de son groupe et la distance entre deux centres de groupes. Les valeurs varient dans $[0; +\infty]$, avec un score de 0 indiquant la classification la plus optimale.

Ces deux indices de performance sont complémentaires, puisque l'indice de Davies-Boudlin offre une évaluation plus globale des clusters, tandis que l'indice de Silhouette évalue la qualité à partir de chaque donnée individuellement. Le calcul de ces deux indices permet ainsi de fournir une évaluation plus robuste.

Pour trouver la valeur de k optimale, il est possible d'itérer sur les valeurs de k, par exemple entre 2 et 20, puis de sélectionner la valeur permettant d'optimiser ces métriques de performance. Lorsque la valeur de k optimisant l'un des indices n'est pas la même pour le second indice, je sélectionne la valeur optimisant l'indice de Silhouette, car elle repose sur une évaluation au niveau individuel et fournit une évaluation plus intuitive.

Étape 3 : différenciation des comportements de polarisation

La troisième étape permet d'évaluer l'impact de la transformation polynomiale et de la pondération des facteurs appliquées dans la métrique GRAIL. Pour cette étape, les paramètres suivants de GRAIL doivent être optimisés :

— Le paramètre a de la transformation polynomiale, influençant la forme et la pente de la courbe sigmoïde, et donc la façon dont les termes basés sur l'entropie sont transformés (Équation (2.3)).

— Le paramètre α, pondérant les facteurs combinés dans GRAIL (Équation (2.6)).

Pour cette optimisation, le paramètre α varie entre 0 et 1 avec $\Delta_\alpha = 0,1$. De plus, a varie de façon à prendre des valeurs inférieures, égales ou supérieures à 1 pour faire varier la forme et la pente de la courbe de la fonction polynomiale (Équation (2.3)). Je choisis ainsi de faire varier a selon les valeurs suivantes : $[1/4; 1/3; 1/2; 1; 2; 3; 4]$. Au total, 77 combinaisons de paramètres sont testées. Une fois les facteurs calculés pour chacun des utilisateurs, puis transformés et pondérés selon une certaine combinaison de paramètres, l'algorithme de clustering k-means est appliqué et évalué selon la méthodologie présentée dans l'étape 2 de ce protocole. Les paramètres permettant d'obtenir les performances de clustering les plus élevées sont ainsi sélectionnés et appliqués.

Étape 4 : pertinence des valeurs de GRAIL

Les conclusions qui peuvent être tirées des étapes d'évaluation précédentes se limitent à valider la modélisation des facteurs de polarisation proposée par GRAIL, et ne permettent donc pas de conclure sur la pertinence de la combinaison de ces facteurs en une valeur unique pour quantifier la polarisation de façon individuelle et multi-factorielle. L'objectif de cette dernière étape est ainsi d'analyser dans quelle mesure GRAIL fournit une quantification fiable de la polarisation, participant à la modélisation du phénomène. Cependant, les valeurs de GRAIL ne peuvent pas être comparées à des valeurs réelles de polarisation, qui ne sont pas disponibles. Pour répondre à cette limite majeure, je propose une méthodologie d'évaluation qui s'appuie sur

l'hypothèse suivante : si la variance des scores de GRAIL dans chaque classe de comportement de polarisation peut être expliquée par des indicateurs comportementaux supplémentaires (*c.-à-d.* non utilisés par GRAIL), cela confirme la capacité de GRAIL à quantifier la polarisation résultant de différents comportements.

Cette méthodologie d'évaluation consiste ainsi à appliquer une régression hiérarchique, qui est un modèle statistique permettant d'analyser la relation entre une variable dépendante et une ou plusieurs variables indépendantes [Arrègle, 2003]. Concrètement, la régression hiérarchique fournit un coefficient de détermination R^2 représentant la proportion de la variabilité de la variable dépendante qui peut être expliquée par une combinaison de variables indépendantes, chacune permettant d'augmenter cette proportion (ΔR^2). Ainsi, plus la valeur de R^2 est élevée, plus les variables indépendantes permettent d'expliquer la variable dépendante. De plus, la relation entre les variables dépendantes et indépendantes est quantifiée au travers d'un coefficient, notés β, informant de l'effet de la variation d'une unité d'une variable indépendante sur la variable dépendante. Une valeur de significativité statistique (p_{valeur}) est également calculée pour évaluer si l'effet de la variable indépendante est statistiquement significatif sur la variable dépendante.

Dans mon contexte d'application, cette régression hiérarchique est ainsi appliquée pour comprendre la relation entre les valeurs de GRAIL (*c.-à-d.* la variable dépendante) dans chaque classe de comportement identifiée, et certains indicateurs comportementaux liés à la polarisation (*c.-à-d.* les variables indépendantes). Ces indicateurs de comportement sont calculés à partir des données étudiées, et sont différents des facteurs de polarisation évalués au travers de GRAIL, apportant des informations quant au comportement de chacun des utilisateurs. Si la variance des valeurs de GRAIL dans chacune des classes de comportement identifiée peut-être expliquée par des indicateurs de comportement, cela confirme la pertinence de la métrique pour quantifier le phénomène de polarisation.

2.4 Évaluation expérimentale

Je détaille premièrement dans cette section les données sur lesquelles la métrique GRAIL est évaluée. Ensuite, la configuration expérimentale nécessaire à la mise en place du protocole d'évaluation présenté sur ces données est détaillée. Finalement, les différents résultats obtenus lors de l'application du protocole expérimental son présentés.

2.4.1 Contexte applicatif : le réseau social Twitter

L'ensemble des travaux présentés dans ce premier chapitre sont appliqués au contexte particulier des réseaux sociaux, catalyseurs du phénomène de polarisation et reflétant fidèlement les comportements adoptés quotidiennement par les utilisateurs. Le contexte d'étude reste le même pour l'ensemble de ce chapitre, ainsi que le suivant.

Pour ce travail sur la modélisation utilisateur, j'ai fait le choix d'étudier une **population d'utilisateurs du réseau social Twitter**, devenu X, qui est particulièrement intéressante du fait de sa sensibilisation accrue aux questions politiques [Boyadjian *et al.*, 2014, Walker and Matsa, 2021]. Cette population a un niveau élevé d'intérêt politique, et il est possible d'identifier l'orientation idéologique des utilisateurs sur la base des positions prises sur certaines questions dans le débat public [Barberá, 2020]. Par ailleurs, les discussions sur Twitter reflètent principalement les préoccupations et les sujets abordés par les médias grand public. À certains égards, ce réseau social Twitter semble étroitement corrélé au cadrage médiatique [Russell Neuman *et al.*, 2014, McCombs and Shaw, 1972]. Les utilisateurs de Twitter constituent également une population de leaders d'opinion [Katz *et al.*, 2017] qui sont particulièrement sou-

mis aux effets puissants de l'environnement et de l'exposition sélective aux médias. Finalement, afin d'assurer une certaine compréhension du contexte étudié, et pour aller au-delà du traditionnel contexte américain, mes travaux se focalisent sur le contexte français.

Les données relatives au **débat sur le vaccin contre la COVID-19** ont été collectées, ayant suscité un vif intérêt des utilisateurs de Twitter suite à la pandémie, et sur lequel ces utilisateurs se sont forgés une opinion. La collecte des données sur ce débat est basée sur le concept d'utilisateurs élites, défini par [Primario *et al.*, 2017], et désignant des utilisateurs populaires et légitimes pour aborder ce débat (médecins, politiciens, scientifiques, *etc*). Ces utilisateurs élites ont une position connue sur le débat, qu'ils défendent au travers de courts messages appelés *tweets*. Ils représentent ainsi les **sources d'information** partageant des contenus pro- ou anti-vaccins, auxquels sont confrontés les utilisateurs du réseau social Twitter. C'est justement les comportements de ces autres utilisateurs, appelés utilisateurs standards, qui sont confrontés à l'information et la partagent, que je cherche à modéliser. Les *retweets* effectués par ces utilisateurs entre le 1^{er} janvier 2022 et le 31 juillet 2022 ont ainsi été collectés et étudiés. Ces *retweets* correspondent au partage des informations diffusées par les utilisateurs. Pour l'étude du phénomène de polarisation, je pose l'hypothèse que ces *retweets* traduisent une approbation du message publié [Conover *et al.*, 2011]. La méthodologie complète de collecte des données est détaillée dans l'Annexe A de ce manuscrit.

Suivant cette méthodologie, des utilisateurs élites ayant une voix légitime pour s'exprimer sur le sujet du vaccin contre la COVID-19 et dont la position sur le débat est connue, ont été sélectionnés en fonction de leur domaine d'expertise, leur métier et leur popularité sur Twitter. Plus précisément, 10 utilisateurs élites de la communauté pro-vaccins et 10 utilisateurs élites de la communauté anti-vaccins ont été identifiés. La sélection de 10 sources d'information dans chacune des communautés permet de représenter une diversité des sources suffisamment élevée. À partir des *tweets* publiés par ces utilisateurs élites, l'ensemble final des données collectées contient 6 697 *tweets* publiés pendant la période étudiée de 7 mois. Parmi eux, 1 869 *tweets* proviennent d'utilisateurs élites pro-vaccins, tandis que 4 828 proviennent d'utilisateurs élites anti-vaccins. Parmi tous les utilisateurs standards ayant retweeté ces *tweets*, les 500 utilisateurs les plus actifs sur toute la période de collecte dans chacune des communautés ont été étudiés. Cette sélection de 500 utilisateurs actifs dans chaque communauté permet d'assurer l'analyse d'interactions d'utilisateurs intéressés par le débat, et représentatifs de l'ensemble des utilisateurs. Cet ensemble de 1 000 utilisateurs a effectué un total de 16 791 *tweets* sur le contenu pro-vaccins et 283 088 *retweets* sur le contenu anti-vaccins.

À titre exploratoire, j'ai évalué dans quelle mesure cet ensemble de données représente réellement un environnement hautement polarisé. Je représente pour cela les données associées sous la forme d'un graphe présenté dans la Figure 2.6. Les nœuds représentent les utilisateurs, à la fois standards et élites, tandis que chaque arête correspond au *retweet* d'un utilisateur standard sur un *tweet* d'un utilisateur élite. Ce graphe a été construit à l'aide d'un logiciel d'analyse et de visualisation de graphe appelé Gephi [17].

L'analyse visuelle de ce graphe permet en premier lieu d'identifier une distinction claire entre les deux communautés, avec peu d'interactions (*c.-à-d.* d'arêtes) entre elles. En appliquant un algorithme de détection de communauté reposant sur le calcul de modularité, introduit dans l'état de l'art et détaillée dans l'Équation (1.5) [Clauset *et al.*, 2004], ces deux communautés opposées sont en effet identifiées. Une première communauté correspondant aux utilisateurs pro-

17. https://gephi.org/

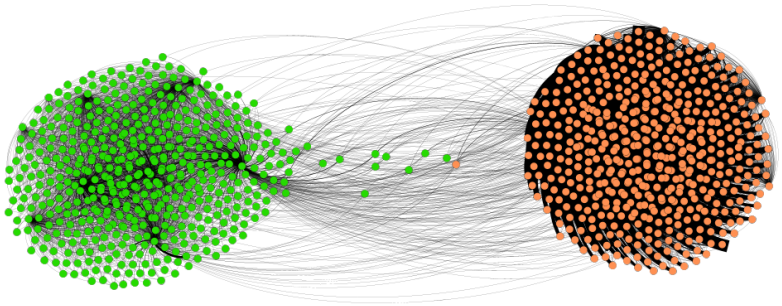


FIGURE 2.6 – Représentation des données collectées sur le débat sur le vaccin contre la COVID-19 sous forme de graphe.

vaccins (en vert dans la Figure 2.6), et une seconde communauté correspondant aux utilisateurs anti-vaccins (en orange dans la Figure 2.6). Cette division du graphe en communautés distinctes est confirmée par la valeur de modularité calculée $(0,55)$.

Par ailleurs, la controversialité du débat sur le vaccin évaluée à partir du processus de marche aléatoire décrit par [Garimella *et al.*, 2018], et présenté dans l'état de l'art, est élevée $(= 0,9)$. Cette forte controversialité indique qu'une fois associé à une communauté, il devient difficile pour un utilisateur d'être confronté à du contenu de la communauté opposée. Cette analyse exploratoire confirme que les données collectées représentent un environnement fortement polarisé. Il existe peu d'utilisateurs intermédiaires, *c.-à-d.* les utilisateurs à l'interface entre les deux communautés. Enfin, les utilisateurs anti-vaccins sont beaucoup plus actifs que les utilisateurs pro-vaccins, avec près de 17 fois plus de retweets dans la communauté anti-vaccins que dans la communauté pro-vaccins.

2.4.2 Étape 0 : Configuration expérimentale

Je détaille maintenant les différents éléments de configuration que j'applique pour l'évaluation de la métrique GRAIL appliquée au jeu de données Twitter collecté.

Facteurs de polarisation étudiés

Pour la validation expérimentale de la métrique GRAIL appliquée au jeu de données Twitter présenté, je choisis d'évaluer les deux facteurs de polarisation, GRAIL est donc bi-factoriel. Les facteurs définis sont les suivants :

— **Facteur opinions**, noté H'_{op}, où une opinion est définie en fonction de la communauté au sein de laquelle un *tweet* a été publié. Ce facteur est alors évalué à partir des *retweets* des utilisateurs standards dans chaque communauté (pro- ou anti-vaccins) ;

— **Facteur sources**, noté H'_{so}, où les utilisateurs élites sont considérés comme les sources d'information sur le sujet du vaccin contre la COVID-19. Ce facteur est alors évalué à partir des *retweets* des utilisateurs standards sur chaque utilisateur élite.

Ces facteurs sont d'autant plus intéressants qu'ils évaluent des facteurs similaires à ceux des métriques baseline sélectionnées. En effet, le score de polarisation ρ est évalué à partir des communautés (*c.-à-d.* des opinions) avec lesquelles un utilisateur interagit, et le manque de

diversité LD est évalué à partir des sources d'information avec lesquelles un utilisateur interagit. La comparaison entre ces baselines et les facteurs GRAIL est donc pertinente.

Pour compléter l'évaluation d'une approche multi-factorielle, et étant donné la répartition équitable des utilisateurs élites (*c.-à-d.* des sources) entre les deux communautés opposées, je propose de diviser le facteur source en deux facteurs afin d'évaluer une version tri-factorielle de GRAIL. Pour cela, le facteur sources initial H'_{so}, qui combine les sources des deux communautés, est divisé en deux facteurs $H'_{so,pro}$ et $H'_{so,anti}$. Le premier correspond au facteur sources calculé pour la communauté pro-vaccins, et le second correspond au facteur sources calculé pour la communauté anti-vaccins. Lorsque trois facteurs sont considérés, et pour que l'évaluation associée soit comparable, je propose d'adapter la métrique baseline LD de la même façon que le facteur sources H'_{so}. Concrètement, la métrique LD est calculée pour les sources de chaque communauté indépendamment, résultant en un premier facteur LD_{pro}, qui correspond à LD évalué sur des sources pro-vaccins, et un second facteur LD_{anti}, qui correspond à LD évalué sur des sources anti-vaccins.

Enfin, je précise que lors de l'optimisation des paramètres de GRAIL, la valeur de α est fixée sur le facteur opinions (ρ ou H'_{op}), et $(1-\alpha)$ sur le facteur sources (LD ou H'_{so}). Lorsque deux facteurs sources sont considérés, $(1-\alpha)$ est réparti uniformément entre LD_{pro} et LD_{anti}, ou $H'_{so,pro}$ et $H'_{so,anti}$.

Orientation de la métrique

Par ailleurs, les données étudiées traitant d'un débat opposant deux communautés, je propose d'orienter le facteur opinion H'_{op}, comme proposé par la métrique GRAIL (Section 2.2.4, Équation (2.7)). Cette orientation du facteur opinions permet d'informer sur la nature de la polarisation des utilisateurs, et donc d'indiquer dans quelle communauté ils sont polarisés. La transformation appliquée est détaillée dans l'Équation (2.9).

$$H^{\pm}_{op} = \frac{\pm H'_{op} + 1}{2} \tag{2.9}$$

Le signe $\pm$ appliqué H'_{op} dépend de la communauté prédominante : je fixe $H'_{op} > 0$ si les interactions sont en faveur de la communauté pro-vaccins, et $H'_{op} < 0$ si les interactions sont en faveur de la communauté anti-vaccins. Les valeurs finales de $H^{\pm}_{op}$ varient dans $[0;1]$, $H^{\pm}_{op} = 0$ indiquant une très forte polarisation dans la communauté anti-vaccins, $H^{\pm}_{op} = 1$, indiquant une polarisation très forte dans la communauté pro-vaccins et $H^{\pm}_{op} = 0,5$ reflétant des interactions équilibrées entre les deux communautés. Ici encore, pour assurer une évaluation pertinente des facteurs de GRAIL, le score de polarisation ρ est également orienté pour pouvoir être comparée au facteur opinions orienté $H^{\pm}_{op}$. De la même façon, ce score de polarisation est $\rho^{\pm} = \frac{\pm\rho+1}{2}$, avec $\rho^{\pm} = 0$ indiquant une forte polarisation dans la communauté anti-vaccins et $\rho^{\pm} = 1$ indiquant une forte polarisation dans la communauté pro-vaccins.

Facteurs comportementaux

Pour appliquer la régression hiérarchique permettant d'évaluer la pertinence de GRAIL pour quantifier la polarisation (Étape 4 du protocole présenté dans la Section 2.3.2), je propose de calculer des indicateurs comportementaux informant sur l'engagement et la régularité des interactions des utilisateurs dans le débat sur le vaccin contre la COVID-19. Selon moi, de tels indicateurs, qui sont différents des facteurs opinions et sources évalués, fournissent des informations sur l'intérêt des utilisateurs pour le débat sur le vaccin contre la COVID-19. Ils reflètent

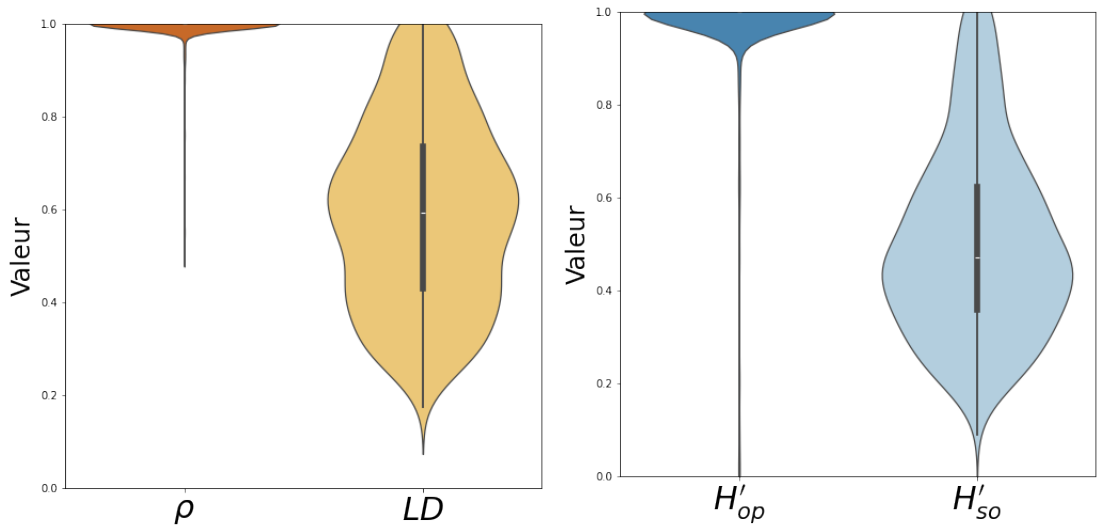
notamment l'implication et l'engagement des utilisateurs dans les discussions abordées sur Twitter. À partir des données collectées, et pour chaque utilisateur standard étudié, les indicateurs suivants sont calculés :

— *NRT* - Nombre de *retweets* concernant le débat sur le vaccin contre la COVID-19 ;
— *%vaccin* - Proportion de *retweets* concernant le débat sur les vaccins par rapport à l'ensemble des *retweets* effectués ;
— *%semaines* - Proportion de semaines actives, *c.-à-d.* le nombre de semaines au cours desquelles l'utilisateur a retweeté ;
— *NPro* - Nombre d'utilisateurs élites de la communauté pro-vaccins retweetés ;
— *NAnti* - Nombre d'utilisateurs élites de la communauté anti-vaccins retweetés.

L'application de la régression hiérarchique sur ces facteurs comportementaux permet donc d'étudier dans quelle mesure chacun contribue à expliquer les variances des valeurs de GRAIL au sein de chaque classe de comportement caractérisée suite à l'application de l'algorithme k-means sur les facteurs de polarisation.

2.4.3 Étape 1 : Modélisation des facteurs de polarisation

Pour débuter la validation expérimentale, je compare premièrement les résultats de la modélisation des facteurs de polarisation lorsqu'elle repose sur les métriques baseline ρ et LD ou sur les facteurs basés sur la mesure d'entropie de la métrique GRAIL. Je rappelle ici que les données exploitées pour le calcul de ρ et du facteur H'_{op} sont les mêmes (*c.-à-d.* les interactions avec les différentes communautés). De la même façon, les données exploitées pour le calcul de LD et H'_{so} sont les mêmes (*c.-à-d.* les interactions avec les différentes sources). Pour comparer ces facteurs, je propose de comparer la distribution des deux facteurs de polarisation de GRAIL, H'_{op} et H'_{so}, avec celle de ρ et LD. Ces distributions sont présentées dans la Figure 2.7.



(a) Distribution de ρ et LD (b) Distribution de H'_{op} et H'_{so}

FIGURE 2.7 – Diagrammes en violon de la distribution des facteurs de polarisation de GRAIL et des métriques baseline.

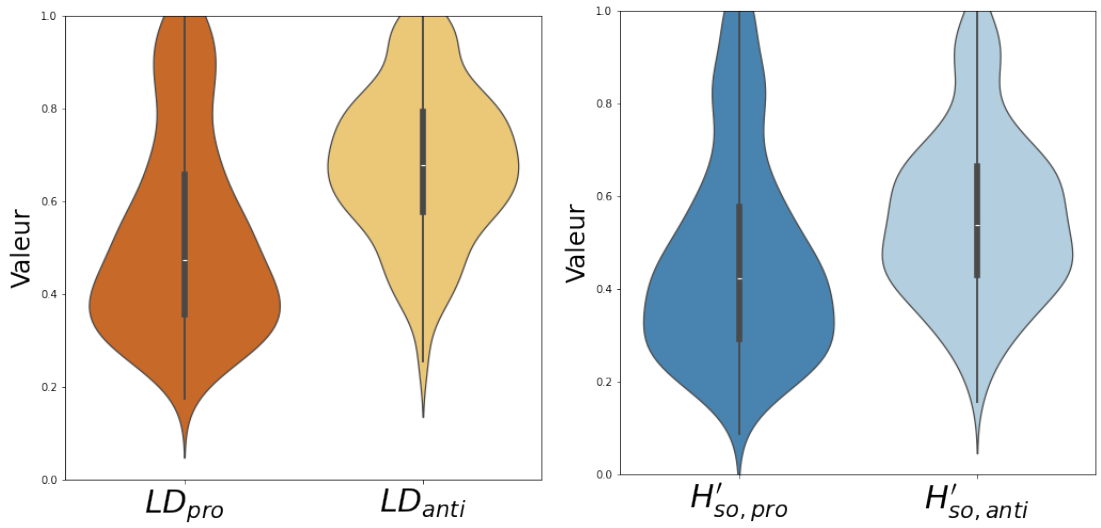
Premièrement, comparant ρ et H'_{op}, les valeurs du facteur opinions H'_{op} sont distribuées dans $[0;1]$ contrairement aux valeurs de ρ. Ce résultat était attendu du fait de la construction de

la métrique ρ, ne considérant que les interactions les plus représentées de chaque utilisateur. C'est pourquoi j'ai proposé de modéliser les facteurs de polarisation à l'aide de l'entropie, afin d'évaluer la diversité des interactions des utilisateurs de façon plus fine. Ce résultat confirme donc que lorsqu'il est modélisé au travers de l'entropie, le facteur opinion H'_{op} est bien distribué dans $[0; 1]$. Cependant, comme une très grande proportion d'utilisateurs n'interagit qu'avec une seule communauté, leur polarisation selon le facteur opinion reste maximale. C'est pourquoi, à l'image du score de polarisation ρ, le facteur opinions H'_{op} reste élevé pour la majorité des utilisateurs, avec $\rho = H'_{op} = 1$. La contribution potentielle du facteur H'_{op} à la différenciation des comportements de polarisation, par rapport à la baseline ρ, concerne donc essentiellement les utilisateurs interagissant dans plus d'une communauté ($H'_{op} \neq 1$).

Deuxièmement, comparant LD et H'_{so}, les distributions diffèrent. En effet, le facteur sources H'_{so} est plus largement distribué que la baseline LD. Cette différence est due, comme pour le score de polarisation ρ, à la façon dont les interactions des utilisateurs sont exploitées pour le calcul de LD, en reposant uniquement sur les interactions majoritaires. Ceci limite ainsi les valeurs à $[0,05; 1]$ lorsque 20 sources sont évaluées, comme c'est le cas ici. Pour approfondir, lorsque les valeurs de LD et H'_{so}, quelle que soit la communauté considérée, sont classées et comparées, le score de corrélation de Spearman obtenu est de 0,92 (p-*value*$= 0$). Bien que relativement élevé, ce score indique qu'une proportion significative d'utilisateurs observe des variations de rang lorsque la diversité de leurs interactions avec les sources est évalué avec LD ou avec H'_{so}. Dans les faits, $49,5$ % des utilisateurs ont une variation supérieure à 5 % des positions. Pour comprendre et expliquer ces variations de rang, un utilisateur, noté u^*, a été sélectionné aléatoirement parmi ceux observant des variations de rang entre LD et H'_{so}. Les *retweets* de l'utilisateur u^* sont, pour plus de la moitié, sur les *tweets* d'un unique utilisateur élite. Les autres *retweets* sont partagés sur 3 autres utilisateurs élites, dont deux pour lesquels il n'y a qu'un *retweet*. Pour cet utilisateur u^*, $LD = 0,52$ et $H'_{so} = 0,91$. La valeur du facteur basé sur l'entropie H'_{so} est donc bien plus élevée et indique un niveau de polarisation élevé par rapport à LD. En tenant à la fois compte du nombre de sources retweetées, mais aussi de la quantité de *retweets* effectués pour chacune de ces sources, le facteur H'_{so} permet donc une quantification plus pertinente de la polarisation par rapport à LD. Pour une proportion significative d'utilisateurs, l'information renvoyée par H'_{so} est donc bien différente de celle renvoyée par LD.

Pour compléter cette évaluation, je propose de comparer les distributions des facteurs de GRAIL et celles des baselines lorsque 3 facteurs sont modélisés. En particulier, je m'intéresse à la distribution lorsque le facteur sources H'_{so} est divisé en deux facteurs $H'_{so,pro}$ et $H'_{so,anti}$, propres à chaque communauté. Ces facteurs sources sont comparés à leur équivalent de la baseline LD, LD_{pro} et LD_{anti}. Les distributions correspondantes sont présentées dans la Figure 2.8.

La différence de distribution observée préalablement entre LD et H'_{so} est à nouveau remarquée entre LD_{pro},LD_{anti} et $H'_{so,pro}$,$H'_{so,anti}$. Les valeurs des facteurs sources évalués avec l'entropie sont plus équitablement répartis dans $[0; 1]$, avec une valeur moyenne inférieure à celle des valeurs calculées à l'aide de la baseline LD, et ce quelle que soit la communauté considérée. Enfin, la comparaison entre LD_{pro} et LD_{anti} d'une part, et entre $H'_{so,pro}$ et $H'_{so,anti}$ d'autre part, permet d'identifier des différences entre les communautés pro- et anti-vaccins. Quelle que soit la façon dont elle est évaluée, avec LD ou avec l'entropie, la diversité des sources est en moyenne plus élevée dans la communauté pro-vaccins que dans la communauté anti-vaccins. En effet, considérant $H'_{so,pro}$ et $H'_{so,anti}$, les valeurs de $H'_{so,pro}$ sont majoritairement situées autour de $0,35$, tandis que les valeurs de $H'_{so,anti}$ sont majoritairement situées autour de $0,55$, traduisant une diversité moins élevée des interactions en fonction des sources d'information. Ceci signifie que les interactions des utilisateurs sont plus distribuées sur les sources pro-vaccin que sur les sources anti-vaccins.



(a) Distribution de LD_{pro} et LD_{anti}

(b) Distribution de $H'_{so,pro}$ et $H'_{so,anti}$

FIGURE 2.8 – Diagrammes en violon de la distribution des facteurs sources de polarisation de GRAIL et des métriques baseline.

En résumé, cette première étape de validation expérimentale permet de confirmer que **les facteurs GRAIL sont plus largement distribués que les métriques baseline, ce qui confirme la valeur ajoutée de l'entropie pour la modélisation des facteurs de polarisation**.

2.4.4 Étape 2 : Modélisation des comportements de polarisation

Pour compléter la validation expérimentale et faire suite à l'évaluation de la modélisation des facteurs de polarisation, j'évalue la modélisation des comportements de polarisation permise par les métriques baseline (ρ et LD) ou par les facteurs basés sur la mesure d'entropie de la métrique GRAIL (H'_{op} et H'_{so}). Cette évaluation est notamment faite sur un nombre de facteurs croissant, passant de 1 à 3 facteurs, afin d'évaluer l'apport de la considération de multiples facteurs pour la modélisation des comportements de polarisation. Je rappelle que ces comportements de polarisation sont caractérisés à partir de l'approche KDE lorsqu'un seul facteur est considéré, et à partir des résultats de l'algorithme de clustering k-means lorsque deux facteurs ou plus sont considérés.

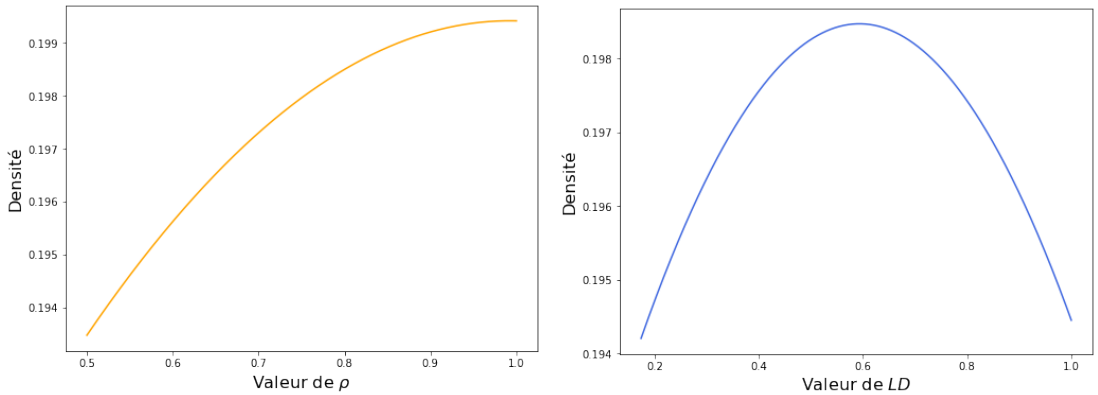
Modélisation uni-factorielle

Pour commencer, j'étudie dans quelle mesure chaque métrique baseline, ρ et LD, permet de distinguer les comportements de polarisation adoptés par les utilisateurs. Cette modélisation uni-factorielle n'est pas appliquée aux facteurs évalués à partir de l'entropie, qui sont approchés de façon multi-factorielle dans la mesure de GRAIL.

Appliquée au score de polarisation ρ, l'application de l'approche KDE ne permet pas d'identifier des maxima ou minima locaux (Figure 2.9a). Cela ne permet pas de différencier des ensembles de valeurs distincts. Encore une fois, c'est l'une des limites préalablement énoncée du score de polarisation ρ (Équation (1.7), page 24), qui est borné dans $[0,5; 1]$ dans un contexte de polarisa-

tion opposant 2 communautés. L'estimation de la densité de probabilité du score de polarisation ρ permet ainsi d'identifier une seule valeur de ρ avec une forte densité, $\rho = 1$. Les autres valeurs de ρ sont donc sous-représentées pour les utilisateurs de l'ensemble de données étudié. La métrique ρ considérée de façon indépendante ne permet donc pas de modéliser des classes de comportement distinctes.

Par ailleurs, l'application de l'approche KDE sur le manque de diversité LD ne présente pas de minima ou de maxima locaux (Figure 2.9b). Seul un maxima global est identifié à $LD = 0,55$, mais il n'existe pas d'autre zone de densité. Au contraire des résultats de l'approche KDE sur ρ, cette absence de zones de densité marquées est étonnante pour LD puisque les valeurs étaient distribuées de façon plus homogène entre les utilisateurs (Figure 2.7). Malgré cette distribution, la modélisation de la polarisation à partir de cette baseline LD seule ne semble donc pas permettre d'identifier des classes de polarisation.



(a) Distribution KDE de ρ (b) Distribution KDE de LD

FIGURE 2.9 – Distributions KDE de ρ et LD.

Cette analyse uni-factorielle introductive permet de mettre en avant que les métriques baseline ne permettent pas de correctement distinguer les utilisateurs entre eux sur le jeu de données étudié, pour lesquelles la majorité des utilisateurs interagissent avec une seule communauté, mais dont la distribution des interactions sur les sources diffère davantage. **Considérée de façon individuelle, ces baselines ne permettent donc pas de différencier des comportements de polarisation.**

Modélisation bi-factorielle

J'étudie si la considération de deux facteurs simultanément participe à une modélisation plus fine des comportements de polarisation. Cette évaluation est tout d'abord appliquée aux métriques baseline ρ et LD, puis aux facteurs opinions et sources de GRAIL, H'_{op} et H'_{so}. Je rappelle que la modélisation des comportements de polarisation est ici abordée au travers d'une approche de clustering, telle que décrite dans la Section 2.3.2. Les performances optimales obtenues à chaque application de l'algorithme k-means pour cette évaluation du protocole, sur différents facteurs, sont récapitulées dans le Tableau 2.2. Pour rappel, les performances sont améliorées lorsque l'indice de Silhouette augmente (valeur optimale = 1), et lorsque l'indice de Davies-Bouldin diminue (valeur optimale = 0).

Facteurs	k **optimal**	**Silhouette** ↗	**Davies-Bouldin** ↘
ρ et LD	2	0,56	0,60
ρ, LD_{pro} et LD_{anti}	2	0,67	0,55
H'_{op} et H'_{so}	3	0,56	0,58
H'_{op}, $H'_{so,pro}$ et $H'_{so,anti}$	4	0,75	0,50

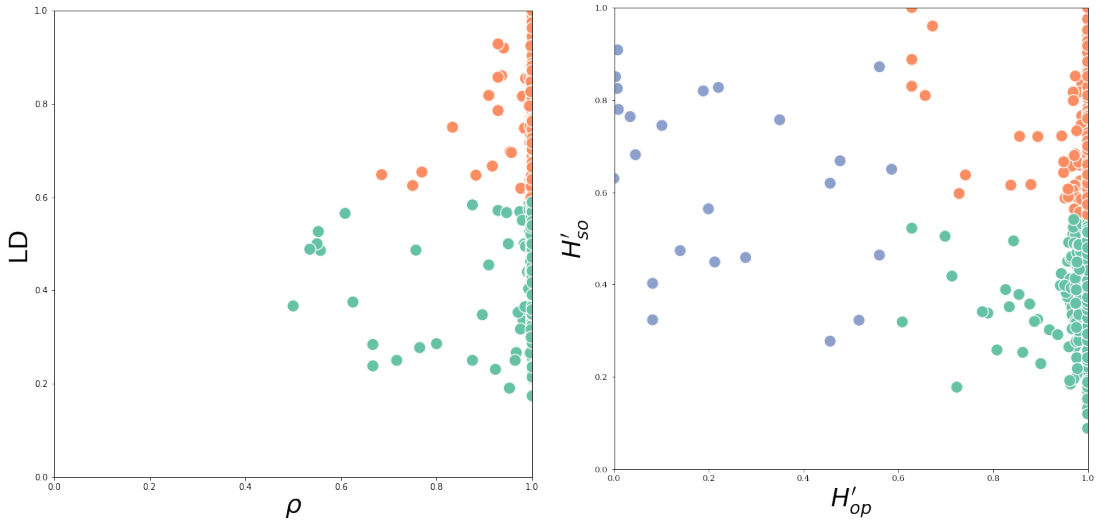
TABLE 2.2 – Performances lorsque l'algorithme k-means est appliqué à différents facteurs

Baselines. Lorsqu'appliqué aux facteurs ρ et LD, les performances de l'algorithme k-means sont optimales avec $k = 2$ (indice de Davies-Bouldin $= 0,60$ et l'indice de Silhouette $= 0,56$). Contrairement à l'analyse à un seul facteur, la prise en compte conjointe des deux métriques permet donc d'identifier deux classes d'utilisateurs adoptant des comportements de polarisation différents (Figure 2.10a). Bien que lorsque considéré comme métrique unique, LD ne permettait pas d'identifier des classes de comportement (Section 2.4.4), il permet ici de différencier les utilisateurs appartenant à chacune des deux classes de comportement identifiées. Un seuil à $LD \approx 0,6$ représente en effet la limite entre les deux clusters. C'est donc l'approche bi-factorielle considérant à la fois ρ et LD qui permet une modélisation plus fine des comportements de polarisation. Ainsi, les utilisateurs avec des valeurs LD plus élevées (cluster orange sur la Figure 2.10a) sont ceux qui ont une forte polarisation selon les sources d'information avec lesquelles ils ont interagi (*c.-à-d.* les utilisateurs élites retweetés), tandis que les utilisateurs avec des valeurs de LD intermédiaires (cluster vert sur la Figure 2.10a) sont ceux qui ont une polarisation plus modérée selon les sources d'information avec lesquelles ils ont interagi.

Toutefois, le seuil à $LD \approx 0,6$ permettant de distinguer les clusters est difficile à interpréter, ce qui appelle à remettre en question la qualité des clusters identifiés. Bien que cette approche bi-factorielle permette de distinguer deux classes de comportement préalablement non identifiés par l'analyse uni-factorielle, elle ne permet pas de caractériser des classes de comportement de polarisation bien distinctes. Selon moi, cette limite est à nouveau liée au calcul de ρ et LD, qui ne tiennent pas compte de l'ensemble des interactions des utilisateurs pour quantifier la polarisation, mais uniquement des interactions les plus représentées.

Entropie. Pour compléter cette évaluation d'une approche bi-factorielle, je propose maintenant d'évaluer dans quelle mesure les facteurs de polarisation évalués à partir de l'entropie, tel que proposé dans le calcul de GRAIL, contribuent à une modélisation plus fine des comportements de polarisation. Ainsi, les facteurs évalués ici sont le facteur opinions H'_{op}, et le facteur sources H'_{so}.

Les performances de l'algorithme k-means appliqué aux données bi-factorielles correspondantes sont optimales avec $k = 3$ (indice de Davies-Bouldin $= 0,58$ et indice de Silhouette $= 0,56$). Les performances sont donc proches de celles obtenues avec ρ et LD, mais un cluster supplémentaire est identifié (Figure 2.10b). Le cluster bleu correspond à un sous-ensemble de 24 utilisateurs (2,4%) dont les valeurs de H'_{op} sont plus faibles, *c.-à-d.* ayant interagi avec les deux communautés (pro-vaccins et anti-vaccins). De manière inédite, les métriques basées sur l'entropie permettent donc de différencier les utilisateurs standards à la fois sur le facteur opinions et le facteur sources, ce qui n'était pas le cas avec ρ et LD. Plus important encore, l'application de k-means sur H'_{op} et H'_{so} contribue à l'identification d'une nouvelle classe de comportement de polarisation intéressante, correspondant aux quelques utilisateurs interagissant dans les deux communautés opposées, qui n'était pas identifiée jusqu'alors.



(a) ρ et LD. (b) H'_{op} et H'_{so}.

FIGURE 2.10 – Clusters identifiés à partir de la modélisation bi-factorielle.

Pour résumer, cette modélisation bi-factorielle permet de confirmer que **la considération de multiples facteurs permet de distinguer des classes de comportement de polarisation**. Par ailleurs, la comparaison entre les facteurs évalués à partir des métriques baseline et ceux de GRAIL basés sur l'entropie permet d'affirmer l'**intérêt de considérer l'ensemble des interactions des utilisateurs pour la modélisation des comportements de polarisation** puisque H'_{op} et H'_{so} permettent de distinguer une troisième classe de comportement intéressante puisqu'elle correspond à des utilisateurs interagissant avec les deux communautés opposées de façon plus équilibrée, et ayant donc une polarisation inférieure aux utilisateurs des autres classes de comportement.

Modélisation tri-factorielle

Pour finaliser cette seconde étape du protocole d'évaluation de GRAIL, je m'intéresse à l'impact d'un troisième facteur sur les résultats de la modélisation des comportements de polarisation. Cette évaluation est tout d'abord appliquée aux métriques baseline $\rho^{\pm}$, LD_{pro} et LD_{anti}, puis aux facteurs opinions et sources de GRAIL, $H^{\pm}_{op}$ et $H'_{so,pro}$ et $H'_{so,anti}$. Pour évaluer des facteurs au plus proche des facteurs finaux constituant la métrique GRAIL, c'est la version orientée du score de polarisation, $\rho^{\pm}$, et du facteur opinions, $H^{\pm}_{op}$ qui est utilisée pour la modélisation (Section 2.4.2).

Baselines. Les performances de l'algorithme k-means appliqué à $\rho^{\pm}$, LD_{pro} et LD_{anti} sont optimales lorsque $k = 2$ (indice de Davies-Bouldin$= 0,55$ et indice de Silhouette$= 0,67$). Les performances de clustering sont donc nettement plus élevées que lors de l'application sur deux facteurs, puisque l'indice de Silhouette est plus élevé ($0,67$ *vs.* $0,56$), et l'indice de Davies-Bouldin est moins élevé ($0,55$ *vs.* $0,60$). Cette amélioration des performances de clustering indique que les données au sein d'un cluster sont mieux regroupées, et que les clusters sont mieux séparés que lors de la modélisation bi-factorielle. L'ajout d'un troisième facteur permet donc de modéliser plus finement les comportements de polarisation.

Plus précisément, les clusters identifiés (Figure 2.11) se différencient par le score de polarisa-

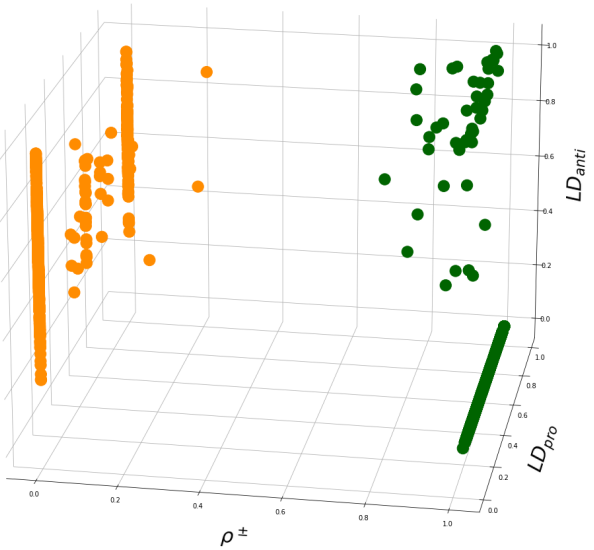


FIGURE 2.11 – Clusters identifiés à partir de la modélisation tri-factorielle avec ρ, LD_{pro} et LD_{anti}.

tion orienté $\rho^{\pm}$, qui ne permettait pas de distinction des utilisateurs lorsqu'il était considéré seul ou avec la métrique baseline LD. Son orientation, couplée à une approche tri-dimensionnelle, permet ainsi de distinguer deux classes de comportements de polarisation, chacune correspondant à des utilisateurs polarisés dans la communauté anti-vaccins (en orange sur la Figure 2.11) ou pro-vaccins (en vert sur la Figure 2.11). Cependant, dans chacun de ces clusters, les métriques baseline LD_{anti} et LD_{pro} ne permettent pas de différencier les utilisateurs. Cette approche tri-factorielle sur les métriques baseline permet donc de mieux différencier les classes de comportements en fonction de la nature de polarisation des utilisateurs (*c.-à-d.* la communauté dans laquelle ils sont polarisés), mais ne permet pas de différencier les utilisateurs en fonction des sources avec lesquelles ils interagissent.

Entropie. Les performances de l'algorithme k-means appliqué à $H^{\pm}_{op}$, $H'_{so,pro}$ et $H'_{so,anti}$ sont optimales lorsque $k = 4$ (indice de Davies-Bouldin$= 0,50$ et indice de Silhouette$= 0,75$). Les performances de clustering sont donc à la fois plus élevées que lors de la modélisation bi-factorielle, et que lors de l'application tri-factorielle sur les métriques baseline. En effet, dans les deux cas, l'indice de Silhouette augmente, et l'indice de Davies-Bouldin diminue. Ces performances de clustering améliorées traduisent l'identification de clusters mieux séparés, et au sein desquels les données sont mieux regroupées. L'approche tri-factorielle reposant sur le facteur opinions $H^{\pm}_{op}$, et les deux facteurs sources $H'_{so,pro}$ et $H'_{so,anti}$, permet donc de mieux modéliser les classes de comportements par rapport à une approche bi-factorielle et par rapport aux baselines.

Les clusters identifiés (Figure 2.12), sont différents de ceux identifiés à partir de la modélisation bi-factorielle (Figure 2.10) ou de la modélisation tri-factorielle basée sur les métriques baseline (Figure 2.11). Tout d'abord, les clusters orange et vert foncé sur la Figure 2.12 cor-

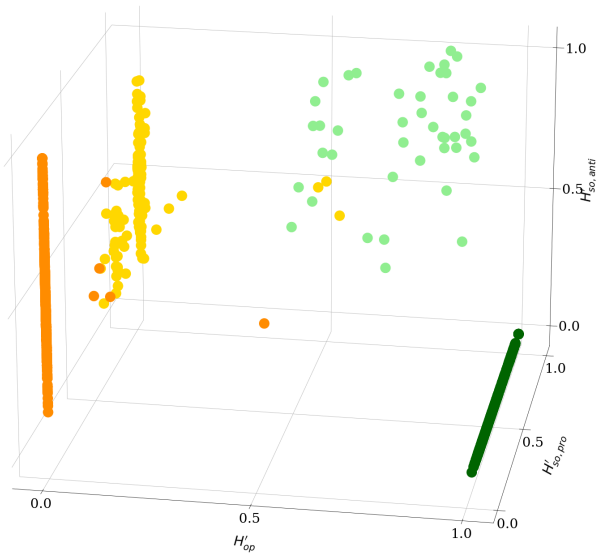


FIGURE 2.12 – Clusters identifiés à partir de la modélisation tri-factorielle ($H^{\pm}_{op}$, $H'_{so,pro}$ et $H'_{so,anti}$)

respondent respectivement à des utilisateurs anti-vaccins et pro-vaccins très polarisés, n'interagissant que dans une seule communauté. Par ailleurs, les clusters de couleur jaune et vert pâle sont particulièrement intéressants. En effet, en examinant plus précisément le facteur opinions $H^{\pm}_{op}$ et les facteurs sources $H'_{so,pro}$ et $H'_{so,anti}$ au sein de ces clusters, les utilisateurs du cluster vert pâle sont ceux dont les valeurs de H'_{op} sont proches de $0,5$, ce qui indique une activité équilibrée entre les deux communautés. Dans chaque communauté, ces utilisateurs interagissent avec diverses sources, les valeurs de $H'_{so,pro}$ et de $H'_{so,anti}$ étant uniformément réparties entre les utilisateurs. En outre, les utilisateurs associés au cluster jaune interagissent principalement avec la communauté anti-vaccins. Néanmoins, ils ont tous au moins une interaction dans la communauté pro-vaccins, dans laquelle ils interagissent principalement avec quelques utilisateurs d'élite ($H'_{so,pro} \approx 1$). Les utilisateurs constituant ces deux clusters sont donc ceux interagissant avec le contenu publié dans les deux communautés, et je choisis de les appeler **utilisateurs intermédiaires**. En plus d'être différents des utilisateurs interagissant au sein des deux communautés identifiées à l'aide de la modélisation bi-factorielle (cluster bleu de la Figure 2.10b), ils sont également plus nombreux puisqu'ils représentent 18,8% des utilisateurs standards (43 utilisateurs pour le cluster vert pâle et 140 pour le cluster jaune), alors que seuls 2,4% des utilisateurs étaient identifiés comme intermédiaires avec l'approche bi-factorielle. Par ailleurs, l'application d'une modélisation tri-factorielle sur les métriques baseline ne permet pas de différencier ces utilisateurs intermédiaires. Seule cette modélisation tri-factorielle basée sur les facteurs évalués avec l'entropie, tel que proposé par la métrique GRAIL, permet donc d'identifier une part significative d'utilisateurs qui interagissent au sein des deux communautés, et adoptent donc des comportements de polarisation moins extrêmes que les utilisateurs ne partageant que le contenu de leur communauté d'appartenance.

Cette modélisation tri-factorielle confirme que la considération de multiples facteurs dans la phase de modélisation conduit à une **modélisation plus fine des comportements de polarisation**. L'identification des clusters associés permet d'avoir une **meilleure compréhension des comportements adoptés au sein du jeu de données étudié**, qui n'est pas permise par une modélisation uni-factorielle, ni à partir des métriques de la littérature qui sont trop restrictives.

2.4.5 Étape 3 : Différentiation des comportements de polarisation

Pour faire suite à l'étape d'évaluation de la modélisation des comportements de polarisation précédente, l'objectif ici est d'évaluer comment les facteurs de polarisation, lorsqu'ils sont transformés avec la transformation polynomiale (Section 2.2.2), puis pondérés dans le calcul final de GRAIL (Section 2.2.3), impactent la discrimination des classes de comportement de polarisation. De façon à fournir une évaluation cohérente des facteurs évalués avec la métrique GRAIL, les facteurs opinions évalués sont les facteurs orientés ($\rho^{\pm}$ et $H^{\pm}_{op}$).

Optimisation des paramètres a et α

Pour cette évaluation, les paramètres a et α relatifs à la transformation polynomiale et à la pondération des facteurs, respectivement, sont optimisés de façon à maximiser les performances de clustering (Section 2.3.2). Les résultats de l'optimisation sont présentés dans le Tableau 2.3.

Facteurs	a **optimal**	α **optimal**	k **optimal**	**Silhouette** ↗	**Davies-Bouldin** ↘
$\rho^{\pm}$, LD_{pro}, LD_{anti}	$a = 1$	$\alpha = 0,5$	2	0,80	0,32
H'_{op}, $H'_{so,pro}$, $H'_{so,anti}$	$a = 1/2$	$\alpha = 0,6$	4	0,85	0,35

TABLE 2.3 – Résultats de l'optimisation des paramètres

D'une part, pour les facteurs $\rho^{\pm}$, LD_{pro} et LD_{anti}, les paramètres optimisés indiquent à la fois que la transformation de ces facteurs ne permet pas de mieux discerner les classes de comportement ($a = 1$, *c.-à-d.* aucune transformation), et que les facteurs opinions et sources ont un poids équivalent ($\alpha = 0,5$). Ainsi calculés, les facteurs permettent d'optimiser les performances du clustering, avec $k = 2$ clusters.

D'autre part, pour les facteurs H'_{op}, $H'_{so,pro}$, et $H'_{so,anti}$, la valeur optimale $a = 1/2$ confirme que la transformation des facteurs basés sur l'entropie suivant la fonction polynomiale de l'Équation(2.3) permet de mieux discriminer les classes de comportement de polarisation. De plus, la valeur optimale de $a < 1$ indique que les utilisateurs interagissent sur un débat très controversé, sur lequel la plupart d'entre eux sont fortement polarisés, tandis que les utilisateurs ayant des scores intermédiaires sont peu nombreux (Section 2.2.2). Par ailleurs, l'optimisation du paramètre α permet de mettre en évidence une légère sur-pondération du facteur opinions H'_{op}, puisque $\alpha = 0,6$ permet d'améliorer les performances du clustering, avec $k = 4$ clusters. Le nombre de classes de comportement identifiées est donc doublé par rapport aux métriques baseline, tout en conservant des performances similaires, ce qui confirme une meilleure distinction des comportements avec ces métriques. Finalement, quels que soient les facteurs considérés, leur transformation et leur pondération permet de considérablement améliorer les performances du clustering par rapport à celui présenté dans la Section 2.4.4 (Tableau 2.2), appliqué sur les facteurs non transformés ni pondérés. Les indices de performance sont en effet maximisés, avec des

indices de Silhouettes augmentés et des indices de Davies-Bouldin diminués, traduisant l'identification de clusters plus homogènes et bien séparés.

Interprétation des classes de comportements identifiées

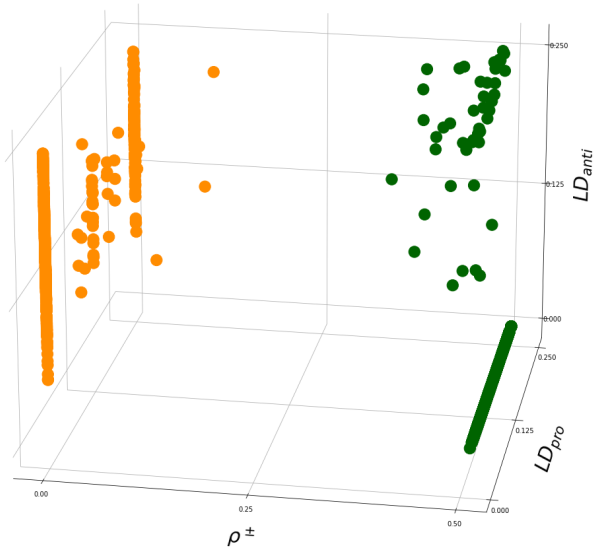
Je propose ici d'analyser et de comparer les clusters identifiés à partir des facteurs $\rho^{\pm}$, LD_{pro} et LD_{anti}, et à partir des facteurs $H^{\pm}_{op}$, $H'_{so,pro}$, et $H'_{so,anti}$ (Figure 2.13).

Je me focalise d'abord sur les deux clusters obtenus avec les facteurs $\rho^{\pm}$, LD_{pro} et LD_{anti} (Figure 2.13a). Ces clusters sont similaires à ceux identifiés suite à la modélisation tri-factorielle (Figure 2.11), sans transformation ni pondération. Les classes identifiées ne se différencient que par le facteur opinions $\rho^{\pm}$. Un premier cluster correspond aux 504 utilisateurs qui accèdent principalement à la communauté anti-vaccins, et le second correspond aux 496 utilisateurs qui accèdent principalement à la communauté pro-vaccins. En raison de l'hétérogénéité des comportements adoptés par les utilisateurs composant ces deux clusters sur les métriques $LD_{so,pro}$ et $LD_{so,anti}$, il est difficile de les caractériser plus précisément. Les utilisateurs sont uniquement répartis en fonction de leur appartenance à l'une ou l'autre communauté, la diversité des sources avec lesquelles ils interagissent n'aidant pas à les discriminer.

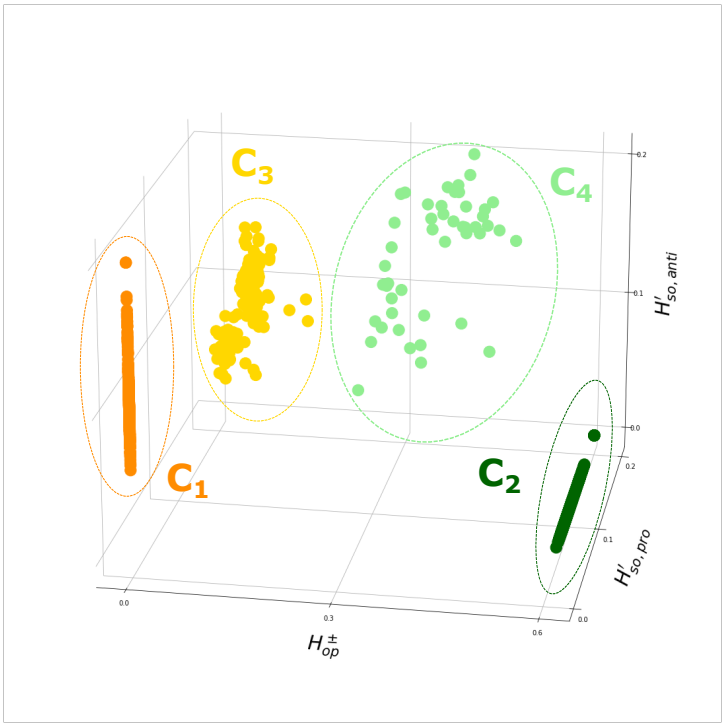
Je me focalise maintenant sur les quatre clusters discriminés sur la base des facteurs $H^{\pm}_{op}$, $H'_{so,pro}$, et $H'_{so,anti}$ (Figure 2.13b). Les clusters sont mieux discriminés puisque les performances de clustering sont améliorées, et les comportements de polarisation adoptés par les utilisateurs dans chacune des classes peuvent ainsi être plus finement caractérisés :

- C_1 (360 utilisateurs) est composé d'utilisateurs très polarisés dans la communauté anti-vaccins. Ils interagissent uniquement au sein de cette communauté et la diversité des sources avec lesquelles ils interagissent est très variable.
- C_2 (452 utilisateurs) est composé d'utilisateurs très polarisés dans la communauté pro-vaccins. Ils interagissent uniquement au sein de cette communauté et la diversité des sources avec lesquelles ils interagissent est très variable.
- C_3 (141 utilisateurs) est composé d'utilisateurs qui interagissent principalement au sein de la communauté anti-vaccins, mais qui ont retweeté au moins un *tweet* de la communauté pro-vaccins. Dans la communauté anti-vaccins, dont ils sont plus proches, les utilisateurs interagissent globalement avec une grande variété de sources, alors qu'ils n'interagissent qu'avec un nombre limité de sources dans la communauté pro-vaccins.
- C_4 (47 utilisateurs) est composé d'utilisateurs qui interagissent dans les deux communautés, avec une préférence pour la communauté pro-vaccins. La répartition de leurs interactions dans chaque communauté est plus équilibrée que pour les utilisateurs du cluster C_3. Dans la communauté pro-vaccins, dont ils sont plus proches, les utilisateurs interagissent globalement avec une variété de sources, alors qu'ils n'interagissent qu'avec un nombre limité de sources dans la communauté anti-vaccins.

Pour résumer, les résultats présentés dans cette section confirment que **les composants de la métrique GRAIL, permettant à la fois de transformer et de pondérer les facteurs de polarisation, contribuent à une meilleure distinction des classes de comportements, quelle que soit la nature des facteurs considérés**. Par ailleurs, **les classes de comportement identifiées à partir des facteurs de GRAIL sont à la fois plus nombreuses et peuvent être caractérisées plus finement, ce qui confirme la pertinence du calcul des facteurs basés sur l'entropie pour la modélisation des comportements**



(a) Clusters identifiés à partir des métriques baseline.



(b) Clusters identifiés à partir des facteurs de GRAIL.

FIGURE 2.13 – Clusters.

de polarisation, qui sont plus adaptés que les métriques individuelles baseline issues de l'état

de l'art.

2.4.6 Étape 4 : Pertinence des valeurs de GRAIL

Les résultats précédents ne permettent pas de conclure sur la fiabilité de GRAIL pour quantifier la polarisation individuelle. Cette quatrième et dernière étape de validation expérimentale y répond en évaluant si les valeurs de cette métrique peuvent être expliquées par des facteurs comportementaux (Section 2.3.2) à l'aide d'une régression hiérarchique. Cette méthodologie est appliquée sur les clusters identifiés lorsque les performances de clustering sont optimales, c'est-à-dire avec une approche tri-factorielle reposant sur le facteur opinions $H\pm_{op}$ et les facteurs sources $H'_{so,pro}$ et $H'_{so,anti}$ (Figure 2.13b).

Compte-tenu des tailles variables des clusters identifiés (entre 47 et 452 utilisateurs), je propose tout d'abord d'appliquer un test de puissance afin d'estimer la taille de l'échantillon nécessaire pour obtenir des résultats statistiques robustes et assurer que les résultats soient statistiquement significatifs [Bourque *et al.*, 2009]. Pour ce test de puissance, je fixe la taille d'effet (*d de Cohen*), correspondant à une mesure de la différence entre deux groupes de données, avec *d de Cohen*$= 0,5$, et une puissance minimale requise de $0,8$, correspondant à la probabilité que la régression hiérarchique aboutisse à un rejet exact de l'hypothèse nulle. Avec ces valeurs de paramètres, la taille minimale requise de l'échantillon est de 34 afin d'assurer la robustesse des résultats fournis par le modèle de régression. L'ensemble des clusters ayant plus de 34 utilisateurs, la régression hiérarchique est appliquée sur les différents facteurs comportementaux présentés. Les résultats de la régression hiérarchique sont présentés dans le Tableau 2.4. Les valeurs finales de R^2 sont données dans la colonne la plus à droite, ainsi que le ΔR^2 apporté par chaque indicateur et le coefficient associé, dont l'ordre et le nombre ont été optimisés pour maximiser R^2. Je rappelle, pour la suite de l'analyse, que les valeurs de GRAIL varient dans $[-1;1]$, -1 indiquant une polarisation extrême dans la communauté anti-vaccins, 1 indiquant une polarisation extrême dans la communauté pro-vaccins et 0 indiquant une absence de polarisation.

TABLE 2.4 – Combinaison optimale d'indicateurs pour la régression hiérarchique pour chaque cluster (C_1 à C_4) avec les coefficients (*p<0.05, **p<0.01, ***p<0.001) et les valeurs de R^2. Les indicateurs sont le nombres de retweets effectués (NRTs), la proportion de retweets sur le débat étudié (%vaccin), la proportion de semaines actives (%semaines), le nombre de d'utilisateurs élites pro-vaccins retweetés (NPro), et le nombre d'utilisateurs élites anti-vaccins retweetés (NAnti).

Cluster	Combinaison optimisée et ordonnée d'indicateurs					R^2
C_1	NRTs	%vaccin	NAnti			**0,65**
β	-0,01***	-0,10***	0,02***			
ΔR^2	0,01	0,43	0,21			
C_2	NRTs	%vaccin	NPro			**0,81**
β	0,01***	0,16***	-0,03***			
ΔR^2	0,05	0,07	0,74			
C_3	NRTs	%vaccin	%semaines	NPro	NAnti	**0,80**
β	-0,01***	-0,24*	-0,03*	0,07***	0,01***	
ΔR^2	0,01	0,16	0,01	0,6	0,02	
C_4	NRTs	%vaccin	%semaines	NPro	NAnti	**0,61**
β	0,01	0,85	-0,15	0,02***	-0,12***	
ΔR^2	0,01	0,03	0,05	0,11	0,41	

Les résultats de la régression hiérarchique permettent tout d'abord de mettre en évidence que les valeurs de R^2 obtenues sont élevées, entre $0,61$ et $0,81$, indiquant que la variance des valeurs de GRAIL peut être principalement expliquée par les facteurs comportementaux étudiés (Section 2.4.2). Je m'intéresse maintenant au détail de ces indicateurs, en analysant et comparant le nombre et la nature des indicateurs de la combinaison optimale de chacun des clusters. Premièrement, la variance des valeurs de GRAIL dans les clusters d'utilisateurs très polarisés (C_1 et C_2) est expliquée uniquement par le nombre de *retweets*, la proportion de *retweets* effectués sur le débat sur le vaccin contre la COVID-19, et le nombre d'utilisateurs élites retweetés dans la communauté d'appartenance de ces utilisateurs polarisés.

Dans la classe C_1 (utilisateurs anti-vaccins très polarisés), les scores de GRAIL sont expliqués par le nombre d'utilisateurs élites de la communauté anti-vaccins retweetés, avec des valeurs de GRAIL augmentant de $0,02$ lorsqu'un utilisateur élite anti-vaccins supplémentaire est retweeté (sources plus diversifiées).

De la même manière, dans la classe C_2 (utilisateurs très polarisés dans la communauté pro-vaccins), les valeurs de GRAIL diminuent lorsque le nombre d'utilisateurs pro-vaccins retweetés augmente. Dans cette classe C_2, le nombre d'utilisateurs élites pro-vaccins retweetés est d'ailleurs l'indicateur principal permettant d'expliquer la variance des valeurs de GRAIL ($\Delta R^2 = 0,74$).

Pour ces deux clusters, une augmentation de l'activité sur le débat, à la fois en nombre de retweets effectués (NRTs) et de proportion de retweets sur la thématique (%vaccin) entraîne une polarisation plus élevée, avec des valeurs de GRAIL se rapprochant de -1 (polarisation extrême dans la communauté anti-vaccins) pour les utilisateurs de C_1, et se rapprochant de 1 (polarisation extrême dans la communauté pro-vaccins) pour les utilisateurs de C_2. Cette proportion de *retweets* sur la thématique (%vaccin) est d'ailleurs l'indicateur expliquant la majorité de la variance pour le cluster C_1 ($\Delta R^2 = 0,43$). Pour ces utilisateurs polarisés dans la communauté anti-vaccins, l'intérêt porté au débat sur le vaccin contre la COVID-19 est donc un indicateur important du niveau de polarisation.

Pour les clusters d'utilisateurs intermédiaires (C_3 et C_4), les indicateurs permettant d'expliquer la variance des valeurs de GRAIL sont plus nombreux. En effet, en plus du nombre de *retweets* et la proportion de *retweets* effectués sur le débat sur le vaccin contre la COVID-19, la proportion de semaines actives et le nombre d'utilisateurs élites retweetés dans les deux communautés permettent également d'expliquer la variance des valeurs de GRAIL. Plus précisément, pour ces deux clusters C_3 et C_4, le nombre d'utilisateurs retweetés dans la communauté de laquelle ils sont la moins proche, est l'indicateur expliquant la majorité de la variance des valeurs de GRAIL ($\Delta R^2 = 0,6$ pour C_3, $\Delta R^2 = 0,41$ pour C_4) : plus les utilisateurs interagissent avec des sources de la communauté opposée, moins ils sont polarisés. Par ailleurs, pour ces deux clusters, l'intérêt pour le débat sur les vaccins, évalué par la proportion d'interactions sur le sujet (*%vaccin*), permet également d'expliquer les variances des valeurs de GRAIL : les utilisateurs qui interagissent beaucoup sur le débat ont tendance à être plus polarisés. Cependant, l'impact de la fréquence des interactions (*%weeks*) n'est pas le même pour les utilisateurs de C_3 et C_4 : lorsqu'ils interagissent plus régulièrement avec le débat, les utilisateurs intermédiaires de C_3 se polarisent davantage, tandis que les utilisateurs intermédiaires de C_4 sont moins polarisés.

Pour résumer, les résultats de la régression hiérarchique mettent premièrement en avant que les classes de comportements identifiées à partir des facteurs de GRAIL sont bien distinctes et reflètent des comportements de polarisation différents. Deuxièmement, la variance des valeurs de GRAIL sont majoritairement expliquées par des indicateurs propres à chaque classe. Cela confirme tout d'abord **la pertinence de GRAIL pour quantifier la polarisation, de façon**

multi-factorielle et individuelle. Par ailleurs, ces résultats permettent d'**affiner la caractérisation des classes de comportement de polarisation identifiées, et donc d'en avoir une compréhension approfondie**. Notamment, les résultats permettent d'observer que les utilisateurs polarisés (C_1 et C_2) se polarisent davantage lorsqu'ils s'intéressent plus au débat. Pour les utilisateurs intermédiaires (C_3 et C_4), une augmentation de l'intérêt pour le débat entraîne également une augmentation de la polarisation, sauf si les interactions sont faites sur des sources d'information de la communauté de laquelle ils sont les moins proches. Enfin, l'impact de la fréquence des interactions sur la polarisation dépend de la communauté dont l'utilisateur est le plus proche.

2.5 Conclusion et discussion

Dans ce chapitre, j'ai présenté GRAIL, la première métrique de polarisation individuelle et multi-factorielle de la littérature. La validation expérimentale de cette métrique GRAIL appliquée au contexte applicatif des réseaux sociaux, a tout d'abord permis de répondre à la première sous-question de recherche (*QR1.1*).

Comment modéliser et combiner de multiples facteurs dans une métrique individuelle de polarisation ? (*QR1.1*)

Je propose de modéliser les facteurs de polarisation au travers de la mesure d'entropie, permettant de tenir compte de l'ensemble des interactions des utilisateurs. Ces facteurs sont ensuite transformés selon une fonction polynomiale, optimisable, afin de différencier les valeurs d'entropie, et de distinguer les utilisateurs entre eux. Pour la combinaison des facteurs, un modèle additif généralisé permet de les pondérer et de les combiner dans une métrique individuelle de polarisation. La métrique résultante est la métrique proposée GRAIL. Elle se différencie des métriques de la littérature par la quantité de données considérées et la façon dont ces dernières sont exploitées pour quantifier finement la polarisation. L'évaluation expérimentale comparant les métriques individuelles de polarisation de la littérature et cette métrique GRAIL a permis de valider la qualité de chaque constituant de GRAIL pour répondre au besoin de modélisation et de quantification de la polarisation. En particulier, la modélisation multi-factorielle permet d'identifier et de caractériser des classes de comportements non identifiées à partir des métriques de la littérature.

Finalement, la dernière étape du protocole d'évaluation, reposant sur l'application de la régression hiérarchique sur différents indicateurs comportementaux, a permis d'évaluer la pertinence de la métrique proposée, et de répondre à la sous-question de recherche (*QR1.2*).

Une telle combinaison permet-elle d'expliquer les comportements de polarisation adoptés ? (*QR1.2*)

Les résultats ont montré que la métrique GRAIL permet d'expliquer finement les comportements de polarisation adoptés à partir de divers indicateurs comportementaux. Pour chacune des classes de comportements identifiées à partir de la modélisation trifactorielle, une description plus précise des comportements, allant au-delà des facteurs de polarisation évalués, peut être donnée. En plus de contribuer à une modélisation plus fine des comportements, ces résultats confirment la pertinence de la métrique proposée.

Ce premier chapitre permet donc de répondre aux limites identifiées de la modélisation du phénomène de polarisation dans la littérature (Section 1). Avec la proposition de la métrique GRAIL, j'ai pu atteindre mes objectifs de travail consistant à proposer une modélisation individuelle et multi-factorielle du phénomène de polarisation. L'ensemble de ces travaux participe ainsi à répondre à ma première question de recherche (*QR1*). Je m'interroge maintenant sur l'évolution temporelle des comportements de polarisation.

Opportunité pour la recommandation. La compréhension approfondie du phénomène de polarisation permise par la modélisation et la quantification individuelle et multi-factorielle du phénomène de polarisation proposée offre de nouvelles opportunités pour la recommandation. La distinction de classes de comportement permet notamment d'envisager des stratégies de recommandation adaptées à chacune d'elle. Dans une perspective de dépolarisation, une adaptation de l'apport en diversité dans les recommandations en fonction du comportement adopté par les utilisateurs peut notamment être envisagée. Cette personnalisation de la diversification permettrait notamment de contrôler les potentiels impacts négatifs de l'apport en diversité, et contribuer à l'élaboration de recommandations fondées sur la confiance. **Les résultats d'une modélisation individuelle et multi-factorielle de la polarisation permet donc d'informer le développement de systèmes de recommandation visant à réduire la polarisation.** C'est cet aspect de recommandation que j'aborderai dans la Partie II de ce manuscrit.

Chapitre 3

Modélisation temporelle des dynamiques de polarisation en ligne

Sommaire

3.1 Introduction

Ce second chapitre de contributions s'intéresse à la modélisation du phénomène de polarisation considérant la dimension temporelle. L'objectif, énoncé à la fin de l'état de l'art, est de proposer une modélisation temporelle du phénomène de polarisation. Cette démarche est motivée par le manque d'une telle modélisation dans la littérature. En effet, les modèles statiques les plus couramment appliqués permettent de modéliser l'état de la polarisation à un instant précis ou sur une période donnée, mais ne permettent en aucun cas d'en saisir les évolutions temporelles. Or, comme expliqué dans l'état de l'art (Section 1), la polarisation évolue dans un environnement dynamique où le débat public est constamment remodelé et dirigé vers de nouvelles questions sociétales. Selon moi, il est donc essentiel de pouvoir modéliser l'émergence de comportements de polarisation lorsqu'un nouveau débat controversé apparaît, ainsi que les variations temporelles associées une fois le débat installé et discuté depuis longtemps. Cette modélisation complète donc la modélisation fine présentée dans le chapitre précédent, contribuant à une meilleure compréhension du phénomène de polarisation.

Les travaux effectués pour répondre à cet objectif sont en lien avec ma première question de recherche (*QR1*) : *Comment modéliser les comportements de polarisation individuels de façon multi-factorielle pour tenir compte de la complexité du phénomène de polarisation, et de façon temporelle pour rendre compte de la dynamique sous-jacente ?* Pour compléter cette question et répondre à l'objectif de modélisation temporelle, ces sous-questions de recherche sont posées :

> ***Sous-questions de recherche abordées dans ce chapitre :***
> Comment la polarisation et les comportements associés évoluent-ils au cours du temps ? (*QR1.3*)
>
> Comment la polarisation se développe-t-elle lorsqu'un nouveau débat controversé émerge dans le discours public ? (*QR1.4*)

Dans la suite de ce chapitre, l'approche de modélisation temporelle proposée est tout d'abord détaillée. Les résultats de l'évaluation expérimentale sont présentés à la suite.

3.2 Approche de modélisation temporelle

L'objectif de l'approche de modélisation temporelle proposée est de compléter la modélisation individuelle et multi-factorielle de façon à rendre compte des potentielles évolutions et dynamiques du phénomène de polarisation. Pour y répondre, je choisis d'**aborder la polarisation comme un processus évolutif**. L'approche ainsi proposée est résumée dans la Figure 3.1, et détaillée dans les sous-sections suivantes.

3.2.1 Initialisation

Pour évaluer le caractère évolutif du phénomène de polarisation, je propose une approche de modélisation temporelle reposant sur des **fenêtres temporelles**.

L'utilisation de fenêtres temporelles est préférée à l'utilisation de données horodatées uniques (*timestamp*) pour plusieurs raisons. Premièrement, les fenêtres permettent de capturer des tendances sur des périodes plus étendues, qui peuvent être difficilement observées en considérant des points temporels isolés. Deuxièmement, les fenêtres temporelles permettent de réduire le bruit qui peut être introduit par l'adoption de comportement non représentatifs adoptés pour certain *timestamp*. Finalement, l'agrégation des données sur des fenêtres permet une meilleure gestion des données manquantes.

Dans une phase d'initialisation du modèle, les fenêtres temporelles sont ainsi définies par leur **durée** d, correspondant à la longueur de temps couverte par chaque fenêtre (1 heure, 2 jours, 3 semaines, 1 mois, *etc.*) et par leur **recouvrement** r, correspondant à la longueur de temps sur laquelle les fenêtres successives se superposent. La définition d'un recouvrement permet de capturer les évolutions en réduisant les biais liés à de potentielles discontinuités entre les fenêtres successives. Ainsi, plus le recouvrement est important, plus les évolutions sont lissées (moins erratiques).

Pour résumer, le caractère évolutif de la polarisation est étudié à partir de son évolution au fil de fenêtres temporelles. Je propose alors d'approcher la modélisation temporelle selon deux points de vue : **(1) un point de vue qualitatif permettant de modéliser l'évolution des comportements de polarisation adoptés par les utilisateurs**, et **(2) un point de vue**

quantitatif permettant de modéliser les dynamiques de polarisation. Dans les deux cas, la modélisation proposée repose sur les principes clés présentés dans le Chapitre 2, à savoir l'adoption d'une approche multi-factorielle et individuelle.

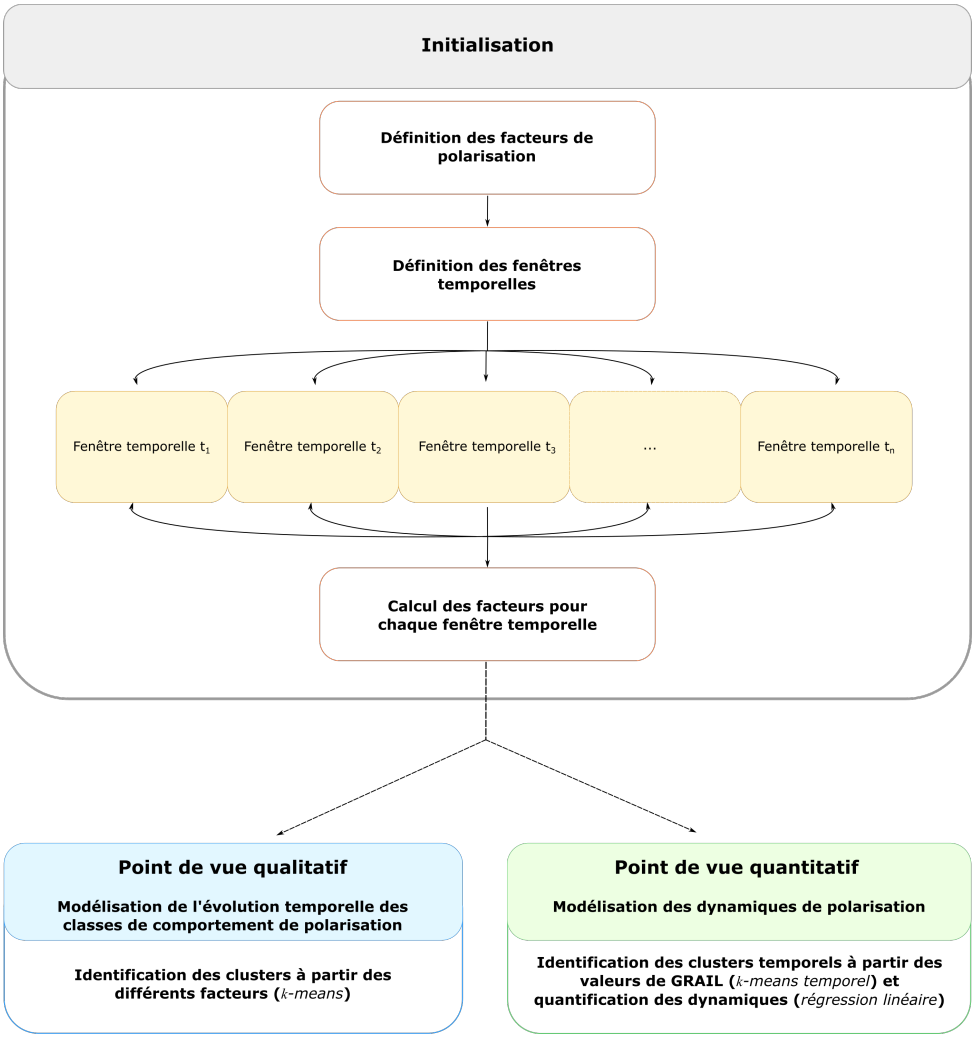


FIGURE 3.1 – Approche de modélisation temporelle du phénomène de polarisation

3.2.2 Point de vue qualitatif : évolution temporelle des comportements de polarisation

Je propose tout d'abord une modélisation temporelle selon un point de vue qualitatif, consistant à étudier l'**évolution temporelle des comportements de polarisation**. Pour cela, différents facteurs de polarisation sont évalués selon la modélisation individuelle et multi-factorielle présentée dans le Chapitre 2 pour chaque fenêtre temporelle. Les classes de comportements sont ensuite identifiées à l'aide de l'algorithme k-means suivant la méthodologie présentée dans la Section 2.3.2. L'application de la modélisation sur des fenêtres temporelles successives permet tout d'abord de **modéliser les variations des classes de comportements identifiées en termes de nombre et de nature**. Les potentielles variations observées peuvent traduire une évolution des comportements de polarisation adoptés par les utilisateurs.

Pour compléter cette étude des variations de classe et tenter de mieux les comprendre, je propose de **caractériser des périodes de polarisation**. Je définis une période comme une séquence de fenêtres temporelles consécutives pour lesquelles les clusters sont à la fois de même nombre, et cohérents. Deux clusters sont considérés comme cohérents s'ils satisfont les conditions de cohérence suivantes :

— Cohérence de la caractérisation : les clusters sont caractérisés de la même façon à partir des facteurs de polarisation qui ont permis de les identifier.

— Cohérence spatiale : les distances entre le cluster considéré et les autres clusters identifiés sont similaires.

Les périodes correspondent ainsi à des périodes temporelles durant lesquelles les comportements de polarisation restent relativement stables.

Ce point de vue qualitatif de l'approche de modélisation temporelle contribue ainsi à une **caractérisation de l'évolution temporelle du phénomène de polarisation**.

3.2.3 Point de vue quantitatif : évolution temporelle de GRAIL

Pour compléter l'approche de modélisation temporelle présentée, je propose d'adopter un point de vue quantitatif consistant à **identifier des dynamiques de polarisation**. Ce ne sont alors plus les classes de comportement qui sont évaluées, mais les valeurs de polarisation GRAIL. Je définis ainsi une **dynamique comme une tendance d'évolution des valeurs de GRAIL, suivies par un ensemble d'utilisateurs**. Ces dynamiques sont identifiées à l'aide de l'algorithme k-means, adapté aux séries temporelles. Le calcul de distance traditionnel appliqué par l'algorithme k-means n'étant pas adapté à l'étude de séries temporelles, une mesure de distance spécifique est appliquée : la déformation temporelle dynamique (*dynamic time warping*) [Sakoe and Chiba, 1978]. Une régression linéaire est ensuite appliquée pour quantifier et caractériser plus précisément les tendances associées aux dynamiques identifiées.

Ce point de vue quantitatif de l'approche de modélisation temporelle contribue ainsi à une **quantification précise de l'évolution temporelle du phénomène de polarisation**.

3.3 Évaluation expérimentale

3.3.1 Contexte applicatif : les débats du vaccin contre la COVID-19 et du conflit en Ukraine sur Twitter

Comme expliqué dans la section 2.4.1, le contexte applicatif de ce chapitre est le même que celui du chapitre précédent, *c.-à-d.* le réseau social Twitter. Néanmoins, pour répondre à la seconde sous-question de recherche (*QR1.4*), des données d'interactions à propos d'un autre débat : **le débat sur le conflit en Ukraine** ont été collectées suivant la méthodologie reposant sur le concept d'utilisateurs élites (Annexe A). Ainsi, comme pour le **débat sur le vaccin contre la COVID-19**, les données ont été collectées entre le 1er janvier 2022 et le 31 juillet 2022. L'intérêt d'avoir collecté des données sur ce second débat est de pouvoir étudier les interactions concernant un débat émergent, puisque le conflit Ukrainien a débuté le 24 février 2022. Ainsi, les deux jeux de données étudiés concernent des **débats de maturité différente**. Une comparaison entre les deux débats est ainsi opérée.

Les données collectées sur le débat en Ukraine représentent 11 205 *tweets* publiés par les 20 utilisateurs élites identifiés (10 utilisateurs pro-Ukraine et 10 utilisateurs pro-Russie) pendant

la période étudiée de 7 mois. Parmi eux, 8 488 tweets ont été rédigés par des utilisateurs élites pro-Russie, et 2 717 tweets proviennent donc d'utilisateurs élites pro-Ukraine. Les interactions des utilisateurs standards étudiés sur ces *tweets* sont à l'origine de 152 802 *retweets*, dont 111 171 sur le contenu pro-Russie et 41 631 sur le contenu pro-Ukraine. Comme effectué pour les données sur le débat du vaccin, et à titre exploratoire, les métriques de modularité et de controversialité ont été calculées à partir du graphe construit à partir des données d'interactions (Figure 3.2). Le score de modularité de $0,38$ indique une division importante des nœuds au sein des communautés, mais moins marquée que pour le débat sur le vaccin contre la COVID-19 ($modularité = 0,55$). Cependant, le score élevé de controversialité calculé $(0,8)$ confirme que le débat oppose deux communautés disjointes. Ainsi, ce débat sur le conflit en Ukraine est fortement polarisé, opposant la communauté pro-Ukraine à la communauté pro-Russie. Par ailleurs, c'est la communauté pro-Russie qui est la plus active.

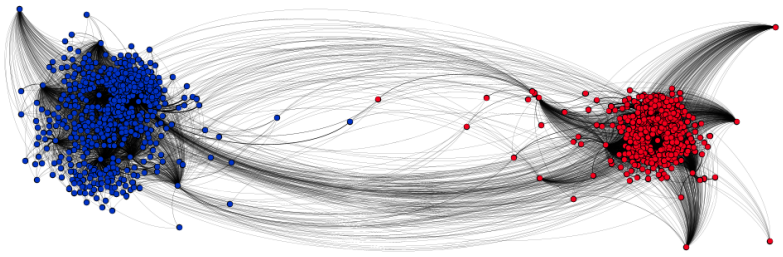


FIGURE 3.2 – Représentation des données collectées sur le débat sur le conflit en Ukraine sous forme de graphe.

Facteurs de polarisation étudiés

Pour la modélisation temporelle des comportements de polarisation appliquée au jeu de données Twitter présenté, je choisis d'évaluer les deux mêmes facteurs de polarisation que dans le Chapitre 2 : le facteurs opinions H'_{op} et le facteur source H'_{so}. Ainsi, les facteurs évalués pour chacun des débats sont le facteur opinions orienté $H\pm_{op}$, et les deux facteurs sources : $H'_{so,pro}$ et $H'_{so,anti}$ pour le débat sur le vaccin contre la COVID-19, et $H'_{so,proU}$ et $H'_{so,proR}$ pour le débat sur le conflit en Ukraine.

Avant d'appliquer la modélisation temporelle des comportements de polarisation, la modélisation individuelle multi-factorielle, telle que présentée dans le Chapitre 2 est appliqué aux données des deux débats étudiés.

Analyse exploratoire des données sur le débat du vaccin contre la COVID-19

Les résultats de cette analyse exploratoire correspondent aux résultats présentés dans le Chapitre 2, Section 2.4.5, page 54.

Analyse exploratoire des données sur le débat du conflit en Ukraine

Les facteurs évalués sont donc le facteur opinions orienté $H^{\pm}_{op}$ (-1 indiquant une polarisation extrême dans la communauté pro-Russie, et 1 indiquant une polarisation extrême dans la com-

munauté pro-Ukraine), le facteur sources dans la communauté pro-Ukraine $H'_{so,proU}$, et le facteur sources dans la communauté pro-Russie $H'_{so,proR}$. Suite à la phase d'optimisation des paramètres de GRAIL, le paramètre a optimal est $1/3$ et le paramètre α optimal est $0,6$. Appliqué aux facteurs transformés et pondérés, l'algorithme k-means a des performances optimales (indice de Silhouette $= 0,87$ et indice de Davies-Bouldin$=0,32$) lorsque $k = 4$ clusters (Figure 3.3).

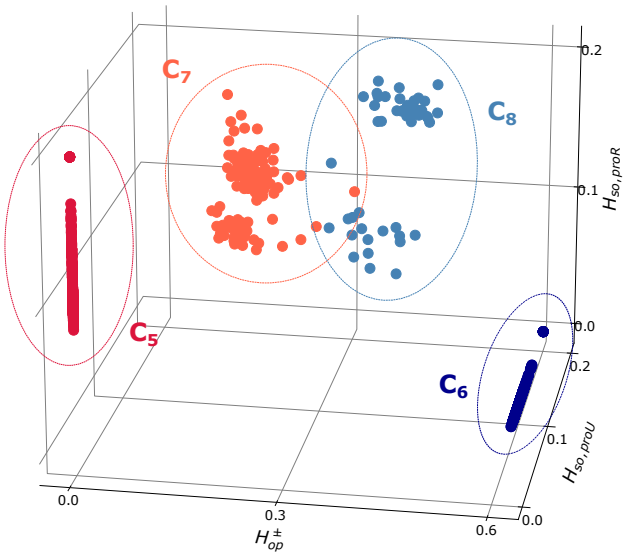


FIGURE 3.3 – Clusters identifiés à partir des facteurs de GRAIL - Conflit en Ukraine

Les valeurs d'entropie modélisant les facteurs de polarisation associées aux utilisateurs composant les clusters identifiés permettent de les caractériser de la façon suivante :

- C_5 (344 utilisateurs) est composé d'utilisateurs très polarisés dans la communauté pro-Russie. Ils interagissent uniquement au sein de cette communauté et la diversité des sources avec lesquelles ils interagissent est très variable.
- C_6 (448 utilisateurs) est composé d'utilisateurs très polarisés dans la communauté pro-Ukraine. Ils interagissent uniquement au sein de cette communauté et la diversité des sources avec lesquelles ils interagissent est très variable.
- C_7 (156 utilisateurs) est composé d'utilisateurs qui interagissent principalement au sein de la communauté pro-Russie, mais qui ont retweeté au moins un *tweet* de la communauté pro-Ukraine. Dans la communauté pro-Russie, dont ils sont plus proches, les utilisateurs interagissent globalement avec une grande variété de sources, alors qu'ils n'interagissent qu'avec un nombre limité de sources dans la communauté pro-Ukraine.
- C_8 (52 utilisateurs) est composé d'utilisateurs qui interagissent dans les deux communautés, avec une préférence pour la communauté pro-Ukraine. La répartition de leurs interactions dans chaque communauté est plus équilibrée que pour les utilisateurs du cluster C_7. Dans la communauté pro-Ukraine, dont ils sont plus proches, les utilisateurs interagissent globalement avec une variété de sources, alors qu'ils n'interagissent qu'avec un nombre limité de sources dans la communauté pro-Russie.

Comparaison des analyses sur les deux débats

Les résultats obtenus suite à l'application de la modélisation individuelle et multi-factorielle pour les deux débats sont comparables. En effet, les clusters identifiés pour le débat sur le conflit en Ukraine sont similaires à ceux identifiés pour le débat sur le vaccin contre la COVID-19 (Figure 2.13b, 56), à la fois en termes de proportions et de caractéristiques. Pour chaque débat, deux clusters d'utilisateurs complètement polarisés et n'interagissant que dans une seule communauté sont identifiés (C_1 et C_2 dans la Figure 2.13b, C_5 et C_6 dans la Figure 3.3). De la même façon, deux clusters d'utilisateurs intermédiaires, interagissant dans les deux communautés mais avec une préférence pour l'une ou l'autre des communautés, sont identifiés pour les deux débats (C_3 et C_4 dans la Figure 2.13b, C_7 et C_8 dans la Figure 3.3).

La modélisation statique des comportements de polarisation adoptés ne permet donc pas de différencier des classes de comportements spécifiques liées à la maturité des débats. Ceci motive davantage la modélisation temporelle pour avoir une compréhension plus précise de l'évolution des comportements de polarisation au cours du temps sur des thématiques émergentes ou installées.

3.3.2 Initialisation

À partir des données collectées, des fenêtres temporelles glissantes d'une durée $d = 4$ semaines, avec un recouvrement $r = 2$ semaines sont définies. Sur les 7 mois de collecte des données, 15 périodes sont ainsi identifiées entre le 1^{er} janvier et le 31 juillet 2022 (Figure 3.4).

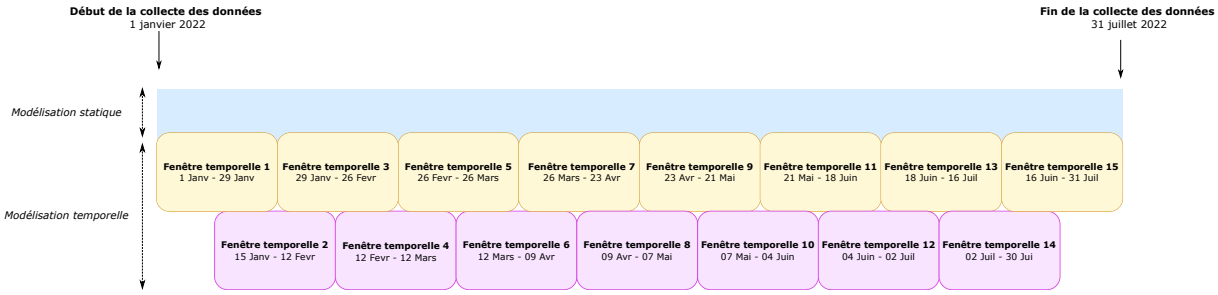


FIGURE 3.4 – Fenêtres temporelles définies pour la modélisation temporelle

Pour la modélisation temporelle des comportements de polarisation, le facteur opinions et les facteurs sources calculés pour chaque utilisateur et pour chacun des deux débats étudiés sont ainsi calculés pour chaque fenêtre temporelle (point de vue qualitatif). De la même façon, les valeurs de GRAIL résultant de la combinaison de ces facteurs sont calculées pour chaque fenêtre temporelle pour la modélisation des dynamiques de polarisation (point de vue quantitatif).

3.3.3 Point de vue qualitatif : évolution temporelle des comportements de polarisation

Les interactions des utilisateurs (*c.-à-d. les retweets*) sont irrégulières durant les 7 mois de collecte, et certains utilisateurs sont inactifs sur certaines fenêtres temporelles. Ainsi, j'ai fixé un seuil de 20 % de périodes inactives, au-dessus duquel les utilisateurs sont écartés de la modélisation temporelle. Parmi les 1 000 utilisateurs étudiés lors de la modélisation statique, 685 et

784 utilisateurs sont suffisamment actifs et sont donc conservés concernant le débat sur le vaccin contre la COVID-19 et le débat sur le conflit en Ukraine, respectivement.

Variation du nombre et de la nature des classes de comportements

Les variations du nombre de classes de comportement identifiées au cours du temps pour chaque débat sont présentées dans la Figure 3.5.

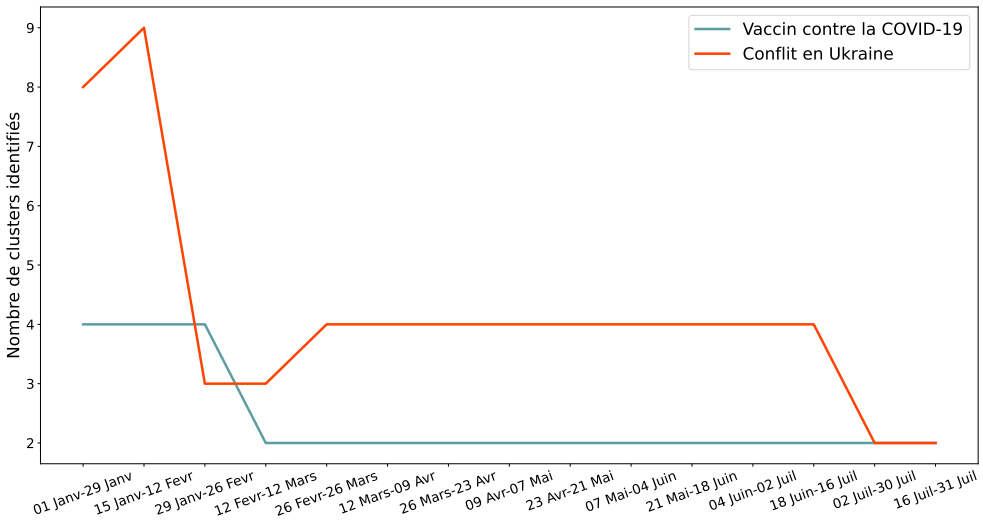
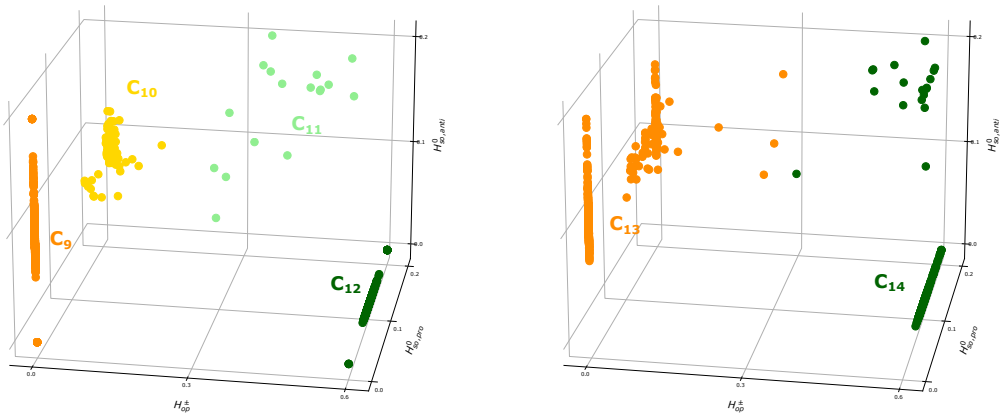


FIGURE 3.5 – Évolution du nombre de clusters identifiés au cours du temps

Tout d'abord, pour le débat de longue durée sur le vaccin contre la COVID-19, le nombre de clusters varie entre 2 clusters au minimum et 4 clusters au maximum. Au cours des trois premières périodes, s'étendant du 1^{er} janvier au 26 février, 4 clusters sont identifiés. Ces clusters (C_9, C_{10}, C_{11} et C_{12} dans la Figure 3.6a) sont cohérents avec ceux identifiés lors de la modélisation statique (C_1, C_2, C_3 et C_4 de la Figure 3.3, Chapitre 2). En effet, deux classes de comportements d'utilisateurs très polarisés, et deux classes de comportements d'utilisateurs intermédiaires sont identifiées. Cependant, dès la quatrième fenêtre temporelle et jusqu'à la fin de la collecte des données, donc du 12 février au 31 juillet 2022, seuls deux clusters sont identifiés (C_{13} et C_{14} dans la Figure 3.6b). Ainsi, dès le mois de février, les classes de comportements d'utilisateurs intermédiaires ne sont plus discernées, et les classes restantes correspondent à des utilisateurs très polarisés, soit dans la communauté anti-vaccins (C_{13}) soit dans la communauté pro-vaccins (C_{14}).

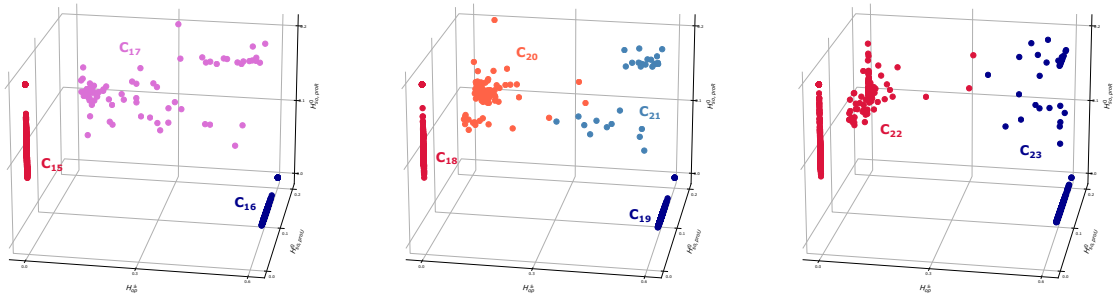
Pour le débat émergent sur le conflit en Ukraine, le nombre de clusters identifiés est beaucoup plus variable, avec 2 clusters au minimum et jusqu'à 9 clusters au maximum. Avant que le conflit ne soit officiellement déclaré (première et deuxième fenêtres temporelles), le nombre de clusters est élevé : 8 et 9 clusters sont identifiés. À partir de la troisième fenêtre temporelle, le nombre de clusters identifiés diminue fortement, avec seulement 3 clusters (C_{15}, C_{16} et C_{17} dans la Figure 3.7a). Plus précisément, deux des trois clusters identifiés (C_{15} et C_{16}) sont semblables aux clusters d'utilisateurs polarisés identifiés lors de la modélisation statique (C_5 et C_6 de la Figure 3.3). Cependant, les deux clusters d'utilisateurs intermédiaires identifiés lors de la modélisation statique (C_7 et C_8 dans la Figure 3.3), ne sont pas discernés, et un seul cluster d'utilisateurs



(a) 4 clusters (b) 2 clusters

FIGURE 3.6 – Clusters identifiées pour le débat sur le vaccin contre la COVID-19.

intermédiaires est identifié (C_{17}). Néanmoins, dès la fin du mois de février, quatre clusters sont identifiés, et ce jusqu'au mois de juillet (C_{18}, C_{19}, C_{20} et C_{21} dans la Figure 3.7b). Au cours de ces fenêtres temporelles consécutives, s'étendant du 26 février au 16 juillet, les clusters sont cohérents avec ceux identifiés lors de la modélisation statique (C_5, C_6, C_7 et C_8 sur la figure 3.3), avec deux classes de comportements correspondant aux utilisateurs polarisés et deux autres classes d'utilisateurs intermédiaires. Enfin, au cours des deux dernières fenêtres temporelles, seuls deux clusters sont distingués (C_{22} et C_{23} dans la figure 3.7c), correspondant à des utilisateurs polarisés dans la communauté pro-Russie (C_{22}) ou pro-Ukraine (C_{23}).



(a) 3 clusters (b) 4 clusters (c) 2 clusters

FIGURE 3.7 – Clusters d'utilisateurs interagissant sur le débat du conflit Ukrainien

Pour résumer, **les clusters identifiés varient à la fois en terme de nombre et de nature au cours du temps**. En effet, **tandis qu'une classe de comportement d'utilisateurs polarisés, dans l'une ou l'autre des communautés, est maintenue, une ou plusieurs classes de comportement d'utilisateurs intermédiaires sont sujettes à des variations**

temporelles. De plus, cette approche temporelle contribue à identifier des différences entre les débats étudiés, dont les classes de comportement associées étaient pourtant semblables lors de l'application d'une modélisation statique. Bien que certaines similitudes puissent être observées entre les deux débats, leur évolution est différente. Ces différences pourraient notamment être associées à leur maturité. Ces résultats renforcent **l'intérêt d'approcher la polarisation comme un processus évolutif, permettant d'identifier des variations temporelles et de souligner des différences liées à la maturité des débats.**

Caractérisation de périodes de polarisation

L'approche de modélisation temporelle permet d'identifier des périodes de polarisation spécifiques, durant lesquelles les clusters identifiés sont de même nombre et cohérents. Suivant la méthodologie proposée, je propose maintenant d'identifier de potentielles périodes de polarisation.

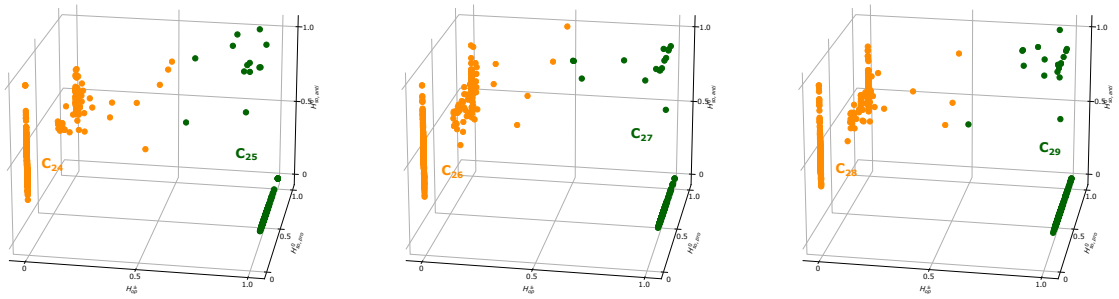
Pour le débat sur le vaccin contre la COVID-19, une première période composée de 4 clusters persiste sur les 3 premières fenêtres temporelles. Elle est suivie d'une seconde période composée de 2 clusters, persistant pendant les 5 mois suivants. Pour le débat sur le conflit en Ukraine, un plus grand nombre de périodes est identifié. La première période d'un mois et demi correspond à la présence de nombreux clusters. Elle est suivie d'une seconde période de même durée, durant laquelle uniquement 3 clusters sont identifiés. Lors de la troisième période, qui est la plus longue sur la période d'analyse des données, un 4*ème* cluster est identifié. Enfin, pour la quatrième et dernière période, la plus courte, seuls 2 clusters sont identifiés.

Suite à l'identification de ces périodes, et afin de comprendre la transition d'une période à une autre résultant de la variation du nombre de clusters identifiés, je propose d'examiner de plus près l'évolution de la distribution des utilisateurs au sein des clusters entre chacune des périodes (Figures 3.6 et 3.7). Les clusters d'utilisateurs polarisés étant stables sur l'ensemble des fenêtres temporelles, je me concentre sur les utilisateurs intermédiaires pour lesquels on observe davantage de variations au cours du temps.

Pour le débat sur le vaccin contre la COVID-19, les utilisateurs intermédiaires de la première période (C_{11} et C_{12} dans la Figure 3.6a), se sont rapprochés des utilisateurs polarisés, pour ne former finalement que deux groupes d'utilisateurs polarisés dans la dernière période (C_{13} et C_{14} dans la figure 3.6b). Pour le débat sur le conflit en Ukraine, les utilisateurs intermédiaires qui n'ont pas de communauté de préférence durant la deuxième période (C_{17} dans la Figure 3.7a) se divisent ensuite en deux clusters d'utilisateurs intermédiaires (C_{20} et C_{21} dans la Figure 3.7b), se rapprochant finalement des utilisateurs polarisés et étant ainsi identifiés comme polarisés durant la dernière période (C_{22} et C_{23} dans la Figure 3.7c).

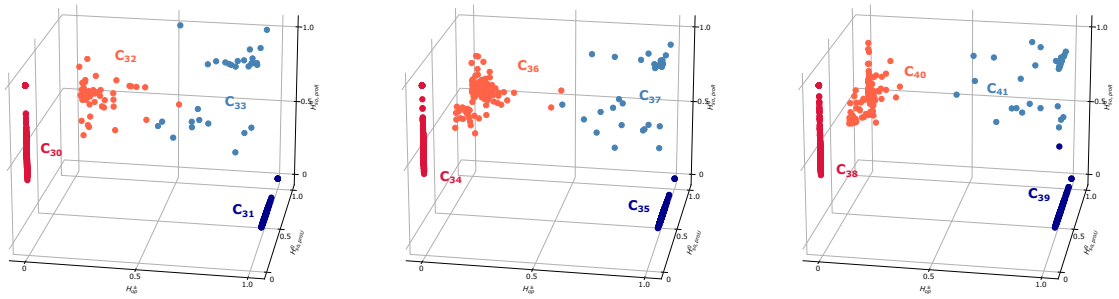
Pour compléter cette analyse inter-périodes, je propose finalement d'analyser l'évolution de la distribution des utilisateurs au cours des périodes les plus longues identifiées pour chaque débat. Considérant tout d'abord l'évolution des utilisateurs au cours de la période de 5 mois pour le débat sur le vaccin COVID-19, s'étendant du 12 février au 31 juillet, les utilisateurs semblent se polariser davantage au fil du temps, en particulier les utilisateurs identifiés comme intermédiaires au cours de la première période (Figure 3.8). Les interactions entre les communautés pro-vaccins et anti-vaccins se raréfient, et cette divergence entre les communautés s'accentue progressivement au cours du temps. En ce qui concerne le débat sur le conflit en Ukraine, l'évolution de la distribution des utilisateurs au cours de la période la plus longue, s'étendant du 26 février au 16 juillet confirme que les utilisateurs intermédiaires se rapprochent progressivement des utili-

sateurs polarisés (Figure 3.9) . Cela reflète une diminution du nombre d'interactions entre les communautés. Lorsque le déséquilibre entre les deux communautés devient trop important, *c.-à-d.* quand les utilisateurs intermédiaires sont trop proches des utilisateurs polarisés, cette période de convergence se termine et seuls deux clusters d'utilisateurs très polarisés sont identifiés.



(a) Position initiale. (b) Position intermédiaire. (c) Position finale.

FIGURE 3.8 – Évolution de la position des utilisateurs ayant interagi avec le débat sur le vaccin COVID-19 pendant la période la plus longue.



(a) Position initiale. (b) Position intermédiaire. (c) Position finale.

FIGURE 3.9 – Évolution de la position des utilisateurs ayant interagi avec le débat sur le conflit Ukrainien pendant la période la plus longue.

L'identification et l'analyse des différentes périodes m'a permis de les caractériser comme suit :

— **Périodes non structurées** : périodes avec de nombreuses classes de comportement, parmi lesquelles aucun comportement spécifique ne peut être identifié.
Dans le débat sur le conflit en Ukraine, elle correspond à la première période. Cette période n'a pas d'équivalent dans le débat sur le vaccin contre la COVID-19.

— **Périodes équilibrées** : périodes comprenant trois classes de comportement : deux classes d'utilisateurs adoptant des positions tranchées (polarisés) et une classe d'utilisateurs maintenant des interactions équilibrées dans les deux communautés opposées.
Cette période est la deuxième identifiée pour le débat sur le conflit en Ukraine, alors qu'elle n'existe pas dans la période analysée pour le débat sur le vaccin contre la COVID-19.

— **Périodes de convergence** : périodes durant lesquelles quatre classes de comportement sont différenciées : deux classes d'utilisateurs polarisés et deux classes d'utilisateurs intermédiaires. Ces derniers se rapprochent des utilisateurs polarisés au cours de la période.
Cette période est identifiée pour les deux débats.

— **Périodes polarisées** : périodes composées de seulement deux classes de comportements, correspondant à des utilisateurs polarisés dans l'une ou l'autre des communautés.
Cette période est identifiée pour les deux débats.

La différence de nombre et de nature des périodes identifiées selon les débats peut s'expliquer par leur maturité, qui influence la façon dont les utilisateurs interagissent sur le sujet. Concernant le débat sur le conflit en Ukraine, la succession des périodes non structurées et équilibrées traduit la transition entre l'absence de comportements de polarisation organisés avant l'émergence du débat (période non structurée) et l'apparition des comportements de polarisation dès la déclaration du conflit (période équilibrée). L'absence de ces deux périodes pour le débat sur le vaccin contre la COVID-19 peut donc s'expliquer par la maturité de ce débat, qui était discuté depuis longtemps au moment de la collecte des données.

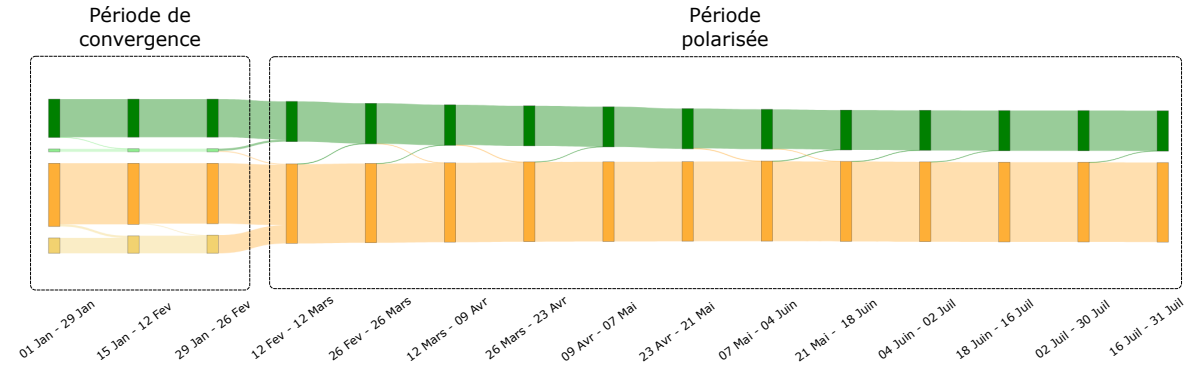
Malgré ces différences, un schéma spécifique de périodes commun aux deux débats se dessine : la période de convergence est systématiquement suivie d'une période de polarisation. Cette succession de périodes reflète un déclin progressif de l'intérêt des utilisateurs intermédiaires pour la communauté qui n'est pas leur communauté principale, et qui finissent systématiquement par se polariser. **L'adoption de comportements de polarisation intermédiaires ne semble donc pas stable dans le temps.** De plus, cette évolution des comportements intermédiaires vers les comportements polarisés peut être plus ou moins rapide puisque la durée des périodes associées varie d'un débat à un autre. Ici encore, les différences peuvent être expliquées par la maturité des débats : plus le débat est émergent, plus les utilisateurs adoptant des comportements intermédiaires mettent du temps à se polariser davantage. La séquence des périodes et l'évolution des clusters associés sont présentés dans Figure 3.10.

Pour résumer, l'analyse de la polarisation par l'intermédiaire de périodes contribue à la fois à mettre en évidence des différences dans l'évolution temporelle des comportements de polarisation liées à la maturité des débats, et des similitudes quant à la succession de périodes spécifiques formant un modèle commun aux deux débats. Ainsi, **l'apparition et l'évolution des classes de comportement de polarisation suit un schéma spécifique de polarisation, dont l'existence et la durée de certaines périodes dépend de la maturité du débat.** La caractérisation des périodes de polarisation, uniquement permise par l'application d'une modélisation temporelle permet donc d'avoir **une meilleure compréhension de la polarisation et de son évolution temporelle.**

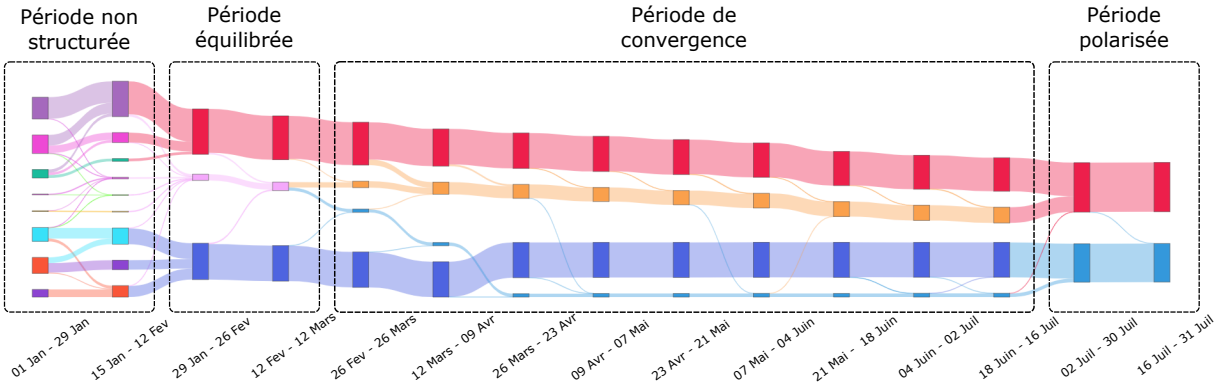
Réflexion sur l'impact du contexte sur l'évolution de la polarisation

Comme détaillé dans l'état de l'art, bien que le phénomène de polarisation soit privilégié par l'utilisation quotidienne des médias sociaux, les comportements de polarisation adoptés par les utilisateurs s'inscrivent dans un contexte plus large. J'émets ici l'hypothèse selon laquelle les différences soulignées entre les débats sont en partie le résultat d'éléments de contexte externes aux réseaux sociaux.

Pour tenter de valider cette hypothèse, je me focalise tout d'abord sur la période non structurée uniquement identifiée pour le débat sur le conflit en Ukraine. Même si les tensions entre la Russie et l'Ukraine pré-existaient et étaient discutées avant février 2022, le cadrage médiatique



(a) Débat sur le vaccin contre la COVID-19.



(b) Débat sur le conflit en Ukraine.

FIGURE 3.10 – Diagrammes de Sankey montrant l'évolution des clusters identifiés durant les différentes périodes de polarisation définies.

de ce sujet était secondaire et n'a pas favorisé l'adoption de comportements de polarisation particuliers. Cependant, dès la déclaration du conflit, cette période destructurée s'achève, suivie par une période équilibrée. La position des utilisateurs intermédiaires durant cette période équilibrée est transitoire du fait de l'émergence d'interactions déséquilibrées entre les deux communautés opposées, qui persistent durant une période de convergence prolongée. Cette longue période peut s'expliquer par le fait que les médias traditionnels ont consacré un espace considérable à la couverture de ce débat durant le printemps 2022. Cette dynamique du débat public nourrit les discussions des utilisateurs dans des contextes externes aux médias sociaux (famille, amis, collègues, *etc.*). Cependant, cette motivation au débat ne change pas nécessairement les idées et convictions déjà partagées [Lazarsfeld *et al.*, 1968].

En ce qui concerne le débat sur le vaccin contre la COVID-19, l'absence de périodes non structurées et équilibrées ne signifie pas qu'elles n'ont pas eu lieu, mais plutôt qu'elles peuvent avoir eu lieu avant le 1^{er} janvier 2022, donc avant la collecte des données. Dans le cas spécifique de la période non structurée, il est probable qu'elle ait eu lieu bien avant janvier 2022, puisque les discussions sur le débat vaccinal ont ressurgi avec la pandémie, et étaient déjà existantes avant. À cet égard, l'identification d'une période de convergence est cependant surprenante puisque le débat était discuté depuis plusieurs mois et était mature en début d'année 2022. En observant les événements et dispositions annoncées par le gouvernement français concernant la vaccination

contre la COVID-19 avant cette période, il s'avère qu'une mesure a été prise quelques temps avant la collecte du jeu de données avec l'ouverture de la campagne de vaccination aux enfants âgés de 5 à 11 ans le 21 décembre 2021. Cette décision a été largement discutée dans les médias et a suscité de vives réactions sur Twitter, rendant certains utilisateurs plus enclins à être confrontés à du contenu partagé dans la communauté opposée. L'existence d'une période d'équilibre à la fin du mois de décembre reste donc possible. Dans tous les cas, la période de convergence a commencé au plus tard le 21 décembre, elle aurait duré au maximum 1,5 mois, ce qui est significativement plus court que la durée de la période de convergence observée pour le débat sur le conflit en Ukraine (5 mois).

D'un point de vue général, cette analyse permet de mettre en avant que, quel que soit le débat, l'émergence d'utilisateurs intermédiaires (périodes d'équilibre ou de convergence) est déclenchée par un **événement perturbateur.** Dans les données étudiées, il y a d'une part, l'invasion de l'Ukraine par la Russie, qui est un événement soudain à propos d'un débat immature ayant suscité de nombreuses interactions, et d'autre part, une décision gouvernementale liée à la campagne vaccinale contre la COVID-19, qui n'est pas soudaine (une annonce a été faite quelques jours avant sa mise en œuvre) et qui intervient dans un débat bien établi. Ces événements sont soutenus par un cadrage médiatique modifié et une augmentation de l'activité des utilisateurs de Twitter sur le débat en question. Cependant, l'impact des deux événements perturbateurs diffère : pour le débat émergent, il conduit d'abord à une période équilibrée, durant laquelle une proportion significative d'utilisateurs ne montre aucune préférence pour l'une ou l'autre communauté, suivie d'une période de convergence qui dure plusieurs mois. Dans le débat mature, il incite certains utilisateurs à rechercher des informations dans la communauté adverse pendant une courte période de convergence.

Pour résumer, cette analyse contextuelle a montré que, quelle que soit la maturité du débat, **des événements perturbateurs peuvent impacter l'adoption de comportements de polarisation spécifiques**, dont la durée dépend à la fois de la maturité du débat et de la soudaineté de l'événement. Ceci permet donc de confirmer l'hypothèse émise en début de section : les différences soulignées entre les débats sont en partie le résultat d'éléments de contexte externes aux réseaux sociaux. Cet effet contextuel permet par ailleurs de souligner que certains utilisateurs adoptant des positions intermédiaires existent sur les débats les plus controversés, et sont ainsi confrontés au contenu publié par la communauté adverse. Cette intermédiarité reste cependant transitoire, et la durée de la période de convergence dépend du degré de polarisation : plus une opinion est cristallisée (*c.-à-d.* une polarisation élevée), plus la convergence vers la communauté d'origine est rapide.

Relation entre les thématiques controversées et les comportements de polarisation associés

Partant du concept d'alignement des enjeux évoqué par [Baumann *et al.*, 2021] (Voir Chapitre 1), je propose maintenant d'étudier la relation entre le débat sur le vaccin contre la COVID-19 et le débat sur le conflit en Ukraine. En effet, les individus appartenant à une communauté spécifique sur un débat appartiennent souvent à une communauté prévisible et spécifique sur un autre débat n'ayant aucun lien apparent [Druckman *et al.*, 2013]. Cet alignement des enjeux est un défi important lié au phénomène de polarisation puisque les individus se polarisent dans des camps similaires selon les mêmes lignes d'idées à propos de différents débats, mais sans nécessairement être correctement informés des arguments sous-jacents.

Pour étudier cette relation, je m'intéresse au comportement du sous-ensemble de 170 utili-

sateurs ayant interagi à la fois sur le débat du vaccin contre la COVID-19 et sur le débat du conflit en Ukraine. Tout d'abord, à partir des résultats de la modélisation statique, je peux mettre en évidence que la communauté à laquelle les utilisateurs appartiennent sur l'un des débats est corrélée avec la communauté à laquelle ils appartiennent sur l'autre débat. En effet, le degré d'association entre les communautés d'appartenance sur les deux débats, évalué à l'aide du V de Cramer, est élevé (V de Cramer = 0,96). Plus précisément, les 133 utilisateurs qui interagissent davantage avec la communauté anti-vaccins interagissent également avec le contenu pro-Russie, tandis que parmi les 37 utilisateurs qui interagissent avec la communauté pro-vaccins, 36 interagissent avec les contenus pro-Ukraine, et un seul utilisateur interagit avec des sources pro-Russie. Ces résultats confirment que, même sur des débats vraisemblablement sans rapport, **l'orientation de la polarisation sur un débat peut être influencée par la position sur un autre débat.** Toutefois, les résultats sont plus modérés quant au degré de la polarisation des utilisateurs : un utilisateur très polarisé sur un débat peut être intermédiaire sur l'autre, et vice versa (V de Cramer = 0,60). Par exemple, dans l'ensemble des 170 utilisateurs interagissant avec les deux communautés, parmi les 75 utilisateurs très polarisés dans la communauté anti-vaccins (C_1 sur la Figure 2.13b), 28% sont intermédiaires sur le débat émergent sur l'Ukraine (C_7 sur la Figure 3.3). Inversement, parmi les 58 utilisateurs intermédiaires qui sont plus proches de la communauté anti-vaccins (C_3 sur la Figure 2.13b), 50% sont également intermédiaires sur le débat sur le conflit en Ukraine, et 50% sont très polarisés. Ceci permet donc de conclure que **la force de la polarisation sur un débat individuel n'a pas nécessairement d'impact sur la force de la polarisation sur d'autres débats, elle influence plutôt l'appartenance à une communauté spécifique**. Les utilisateurs prennent donc position sur les débats émergents selon des lignes d'idées préexistantes, ce qui **confirme l'alignement des enjeux et contribue, une nouvelle fois, à une meilleure compréhension du phénomène de polarisation.**

3.3.4 Point de vue quantitatif : évolution temporelle de GRAIL

Les résultats qualitatifs présentés précédemment ont notamment permis de montrer que les utilisateurs très polarisés observent peu de variations entre janvier et juillet 2022. Pour la modélisation des dynamiques de polarisation, reposant sur la métrique GRAIL et dont l'approche est détaillée dans la Section 3.2, je me focalise donc sur les utilisateurs appartenant aux clusters intermédiaires pour les deux débats. L'identification de dynamiques parmi ces utilisateurs intermédiaires peut participer à une compréhension plus approfondie du phénomène de polarisation.

Débat sur le vaccin contre la COVID-19

Je commence par m'intéresser à l'évolution des valeurs de GRAIL pour les 188 utilisateurs intermédiaires identifiés pendant la période de convergence du débat sur le vaccin contre la COVID-19 (Figure 3.11a). Parmi ces évolutions des valeurs de GRAIL, trois dynamiques de polarisation sont identifiées (indice de Silhouette = 0,96). Une dynamique majeure, en orange sur la Figure 3.11a, représente 86% des utilisateurs intermédiaires et correspond aux utilisateurs se rapprochant progressivement de la communauté anti-vaccins, et donc pour lesquels les valeurs de GRAIL tendent vers -1. Pour ces utilisateurs, le coefficient de régression linéaire est égal à -0.04 ($R^2 = 0.84$), l'interception étant de -0.43 (courbe orange sur la Figure 3.11b). Ceci confirme que les valeurs GRAIL diminuent au cours du temps et que les utilisateurs se polarisent dans la communauté anti-vaccins.

Par ailleurs, une deuxième dynamique correspond à 10% des utilisateurs intermédiaires, s'approchant de la communauté pro-vaccins, en vert sur la Figure 3.11a. Pour ces utilisateurs, le

coefficient de régression linéaire est de $0,02$ ($R^2 = 0,62$), tandis que l'interception est de $0,64$ (courbre verte sur la Figure 3.11b). Ainsi, les valeurs de polarisation des utilisateurs qui suivent cette dynamique sont de plus en plus élevés au fil du temps, se rapprochant de $GRAIL = 1$, et traduisant une polarisation progressive dans la communauté pro-vaccins.

La comparaison de l'évolution des valeurs de GRAIL entre les utilisateurs anti-vaccins (en jaune dans la Figure 3.11b) et pro-vaccins (en vert dans la Figure 3.11b) permet de mettre en avant que les valeurs absolues de polarisation des utilisateurs anti-vaccins augmentent deux fois plus vite que celles des utilisateurs pro-vaccins, puisque le coefficient de régression linéaire est deux fois plus élevé ($0,04$ *vs.* $0,02$), bien qu'elles soient initialement plus faibles ($0,43$ *vs.* $0,64$). Les utilisateurs intermédiaires de la communauté anti-vaccins sont donc moins polarisés dans leur communauté au début de la période de convergence par rapport aux utilisateurs appartenant à la communauté pro-vaccins, mais se polarisent plus rapidement.

Enfin, une dernière dynamique, sous-représentée, correspond à 4% des utilisateurs intermédiaires (en bleu dans la Figure 3.11a) dont les valeurs de GRAIL varient beaucoup. Ceci traduit une polarisation irrégulière dans une communauté puis dans l'autre. Pour ces utilisateurs, la qualité de la régression linéaire est faible (valeurs de R^2 faibles) et n'apportent pas d'informations utiles. Ces utilisateurs sont sous-représentés dans le jeu de données étudié, puisqu'ils ne représentent que 0,07% de l'ensemble des utilisateurs.

Pour résumer, pour la plus grande majorité d'utilisateurs intermédiaires pendant la période de convergence sur le débat sur le vaccin contre la COVID-19, les valeurs absolues de GRAIL évoluent rapidement et tendent vers des valeurs extrêmes. Cela confirme les conclusions de la section précédente, selon lesquelles **les utilisateurs intermédiaires se polarisent davantage au cours du temps, jusqu'à adopter un comportement reflétant une forte polarisation.**

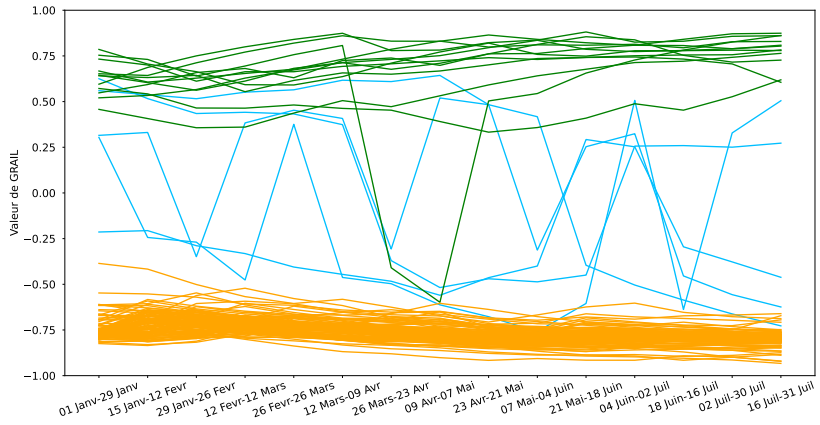
Débat sur le conflit en Ukraine

Je m'intéresse ici à l'évolution des valeurs de GRAIL pour les utilisateurs intermédiaires identifiés pendant la période équilibrée du débat sur le conflit en Ukraine (C_{17} sur la Figure 3.7a), présentées dans la Figure 3.12a. À partir de ces données, 4 dynamiques de polarisation sont identifiées (indice de Silhouette = 0,86). La première, représentée en orange dans la Figure 3.12a, correspond à 58% des utilisateurs et décrit une diminution progressive des valeurs de GRAIL, se rapprochant de -1. Cela confirme que la majorité des utilisateurs intermédiaires se polarisent dans la communauté pro-Russie. Pour ces utilisateurs, le coefficient de régression linéaire est de -0.01 ($R^2 = 0.94$), avec une interception à -0.65 (courbe orange sur la Figure 3.12b).

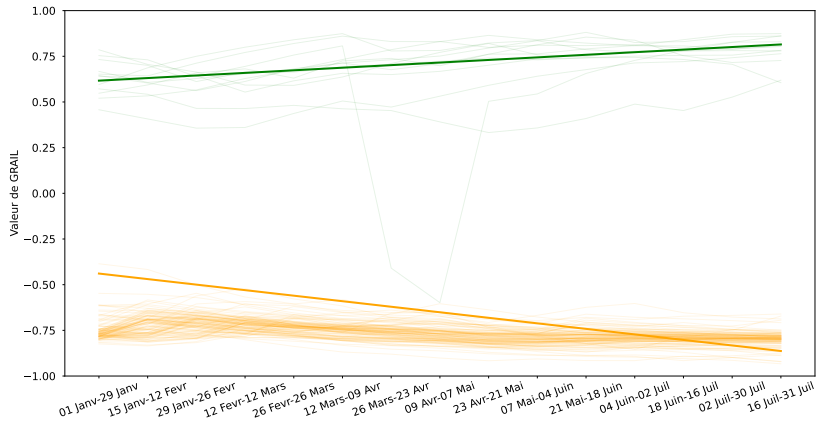
Une dynamique symétrique, correspondant à 26% des utilisateurs, en bleu dans la Figure 3.12a, correspond à une augmentation des valeurs de GRAIL et traduit une polarisation dans la communauté pro-Ukraine. Pour ces utilisateurs, le coefficient est de $0,02$ ($R^2 = 0,96$), et l'interception est de $0,54$ (courbe bleue sur la figure 3.12b).

La comparaison des dynamiques et des évolutions des valeurs de GRAIL permet ainsi de mettre en évidence que les utilisateurs intermédiaires qui rejoignent la communauté pro-Ukraine se polarisent deux fois plus vite que ceux qui rejoignent la communauté pro-Russie, puisque le coefficient de régression linéaire est deux fois plus élevé ($0,02$ *vs.* $0,01$), bien que leurs valeurs absolues de polarisation initiales soient inférieures ($0,54$ *vs.* $0,65$). Les utilisateurs polarisés dans la communauté pro-Ukraine sont donc moins polarisés pendant la période de perturbation, mais se polarisent plus rapidement dès qu'ils entrent dans la période de convergence.

La comparaison des résultats obtenus pour les deux débats étudiés permet de remarquer que les valeurs absolues de GRAIL augmentent deux fois plus vite pour le débat mature sur le vaccin contre la COVID-19 que pour le débat émergent sur le conflit en Ukraine, puisque les coefficients



(a) Dynamiques identifiées parmi les utilisateurs (chaque couleur désigne un cluster temporel)



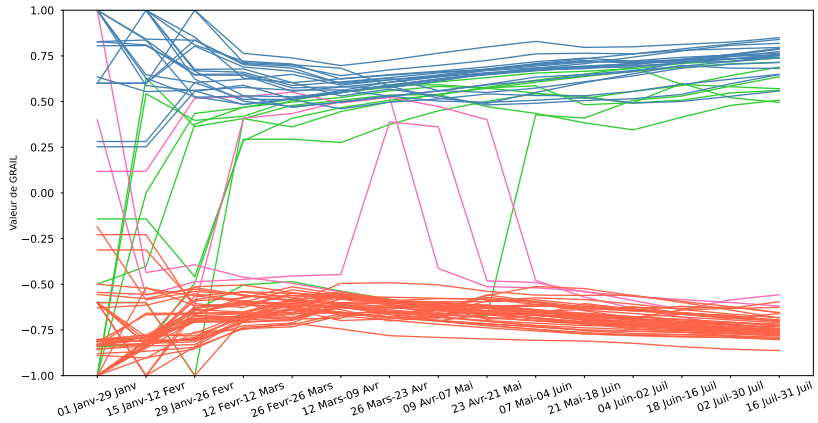
(b) Courbe de relation linéaire entre les valeurs de GRAIL et le temps (chaque couleur désigne la relation linéaire pour un cluster temporel)

FIGURE 3.11 – Évolution temporelle des valeurs de GRAIL pour les utilisateurs intermédiaires sur le débat sur le vaccin contre la COVID-19.

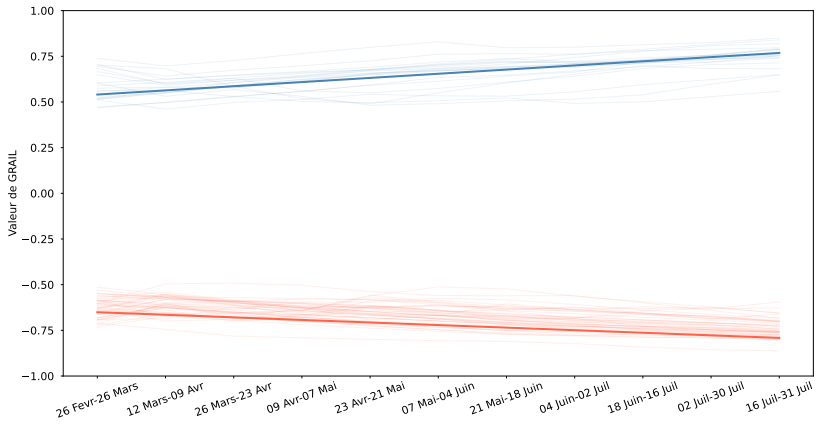
de régression sont plus élevés ($0,02$ et $-0,04$ pour le débat sur le vaccin contre la COVID-19 *vs.* $0,02$ et $-0,01$ pour le débat sur le conflit en Ukraine). Ceci confirme, une nouvelle fois que la maturité du débat impacte l'évolution de la polarisation

Finalement, les deux dynamiques restantes correspondent à des variations extrêmes des valeurs de GRAIL, oscillant entre des valeurs positives et négatives. Plus précisément, 11% des utilisateurs intermédiaires sont initialement plus proches de la communauté pro-Russie et se rapprochent de la communauté pro-Ukraine de façon irrégulière (en vert dans la Figure 3.12a), tandis que 5% d'entre eux observent une variation inverse, de la communauté pro-Ukraine vers la communauté pro-Russie (en rose dans la Figure 3.12a). Pour ces dynamiques, la qualité de la régression linéaire pour les deux dynamiques sous-représentées (en vert et en rose dans la Figure 3.12a) est faible du fait des variations irrégulières observées.

Pour résumer, cette analyse temporelle de l'évolution des valeurs de GRAIL permet de dis-



(a) Dynamiques identifiées parmi les utilisateurs (chaque couleur désigne un cluster temporel)



(b) Courbe de relation linéaire entre les valeurs de GRAIL et le temps (chaque couleur désigne la relation linéaire pour un cluster temporel)

FIGURE 3.12 – Évolution temporelle des valeurs de GRAIL pour les utilisateurs intermédiaires sur le débat sur le conflit en Ukraine.

tinguer différentes dynamiques parmi les utilisateurs intermédiaires à la suite d'un événement perturbateur. En particulier, des différences quant au degré de polarisation suivant l'événement perturbateur et à la vitesse avec laquelle ils se rapprochent de leur communauté et se polarisent ont pu être mises en évidence à la fois entre les communautés opposées, et entre les débats. Ces conclusions viennent confirmer l'**impact de la maturité du débat sur l'adoption de comportements de polarisation particuliers**. Par ailleurs, cette modélisation temporelle basée sur la quantification de la polarisation des utilisateurs, de façon individuelle et multi-factorielle, permet d'affirmer la pertinence de la métrique GRAIL pour la modélisation temporelle.

3.4 Conclusion et discussion

Dans ce chapitre, j'ai présenté une double modélisation temporelle du phénomène de polarisation : une modélisation de l'évolution temporelle des classes de comportements de polarisation (point de vue qualitatif), et une modélisation des dynamiques de polarisation associées (point de vue quantitatif). Ici, la polarisation est envisagée comme un processus évolutif, allant au delà des modélisations statiques les plus répandues dans la littérature. Les résultats présentés permettent de répondre à la sous-question de recherche (*QR1.4*).

Comment la polarisation et les comportements associés évoluent-ils au cours du temps ? (*QR1.4*)

La modélisation temporelle permet de discerner différentes périodes ordonnées caractérisées par l'adoption de comportements de polarisation spécifiques. L'existence de ces périodes est notamment influencée par des événements contextuels, dont le caractère inattendu et l'intensité peuvent avoir des conséquences différentes sur la perturbation et l'évolution de la polarisation.

La comparaison entre le débat sur le vaccin contre la COVID-19 et le débat sur le conflit en Ukraine, dont la maturité au moment de la collecte des données était bien différente, permet de répondre à la sous-question de recherche (*QR1.5*).

Comment la polarisation se développe-t-elle lorsqu'un nouveau débat controversé émerge dans le discours public ? (*QR1.5*)

L'émergence d'un débat controversé entraîne l'apparition d'une période équilibrée de courte durée durant laquelle un sous-ensemble d'individus ne prend pas position. La maturité du débat influence également l'existence et la durée des périodes de polarisation.

Pour compléter, l'étude de la relation entre les deux débats met en évidence l'alignement des enjeux : lorsqu'un nouveau débat émerge, les utilisateurs se polarisent naturellement en suivant les mêmes courants de pensées qu'un débat existant et installé dans le discours public depuis plus longtemps. Cependant, aucune corrélation n'est observée quant au degré de polarisation.

Opportunité pour la recommandation. La compréhension plus complète du phénomène de polarisation permise par l'approche de modélisation temporelle proposée offre de nouvelles opportunités pour la recommandation. L'adoption de comportements de polarisation plus modérés par certains utilisateurs à la suite d'événements perturbateurs prouve à la fois qu'une diminution de l'adoption de comportements extrêmes est possible, mais aussi que la temporalité des recommandations est importante. Par ailleurs, la confirmation de l'alignement des enjeux permet par exemple d'envisager des stratégies de dépolarisation multi-thématiques. **Dans une perspective de dépolarisation au travers de recommandations adaptées, la considération des variations temporelles, des éléments contextuels ainsi que de l'alignement des enjeux peut donc grandement informer les stratégies mises en place**.

Deuxième partie

Apport en diversité dans les recommandations de news

Chapitre 4

État de l'art : systèmes de recommandation et spécificités du contexte des news

Sommaire

Une motivation essentielle de mon travail de thèse est le souhait de développer des approches de diversification personnalisée des recommandations de news afin de répondre au mieux au phénomène de polarisation. Dans cet objectif, et tel que détaillé dans la Partie I de ce manuscrit, le développement d'approches adaptées doit selon moi reposer sur une modélisation fine des comportements de polarisation, qui soit individuelle, multi-factorielle (Chapitre 2) et temporelle (Chapitre 3). Mes intérêts de recherche pour les travaux présentés dans cette Partie II se sont ainsi portés sur la **personnalisation des approches de recommandation**, dont le besoin a notamment pu être mis en évidence par l'identification de classes de comportement distinctes. Le fil conducteur de mon travail réside donc dans l'adoption d'une approche centrée utilisateur, informée par des connaissances pluri-disciplinaires, permettant le développement de modèles de recommandation répondant au phénomène complexe de la polarisation, de façon contrôlée et éthique.

Avant de détailler ma contribution pour cette seconde partie du manuscrit, je présente un état de l'art des systèmes des recommandations et des spécificités du contexte des news.

4.1 Systèmes de recommandation : généralités

A l'aire du *big data*, trouver le contenu d'intérêt parmi les millions de ressources disponibles relève du défi. Les capacités cognitives humaines ne permettent en effet pas de traiter une telle quantité d'information [Golman *et al.*, 2017]. Un filtrage automatisé, et personnalisé, du contenu disponible s'impose alors [Baeza-Yates *et al.*, 1999]. Pour répondre à ce besoin, des approches d'IA, avec notamment la recherche d'information et la recommandation, ont vu le jour [Manning *et al.*, 2008, Resnick and Varian, 1997]. Bien que répondant à un besoin commun de filtrage automatisé et de personnalisation de l'information disponible, ces approches se distinguent par leur mode de fonctionnement : les systèmes de recherche d'information sont centrés sur la sélection d'informations pertinentes en réponse à des requêtes spécifiques exprimées par les utilisateurs (approche *pull*), tandis que les systèmes de recommandation cherchent à anticiper les besoins et les intérêts de ces utilisateurs pour prédire et proposer des contenus susceptibles de leur plaire (approche *push*) [Belkin and Croft, 1992].

La problématique de cette thèse portant sur les systèmes de recommandation, seules ces approches sont détaillées dans la suite de cet état de l'art. Je souhaite tout de même préciser que les défis scientifiques liés à la diversité des informations consommées par les utilisateurs, et le lien avec le phénomène de polarisation (discuté dans la Partie I de ce manuscrit), ne se limitent cependant pas à ces systèmes de recommandation. Les propositions discutées et proposées dans la suite de cette seconde partie du manuscrit sont donc applicables à tous les systèmes de filtrage d'information, y compris les systèmes de recherche d'information.

4.1.1 Principe général de la recommandation

Les systèmes de recommandation ont été conçus pour orienter l'attention des utilisateurs vers des ressources adaptées à leurs besoins ou à leurs préférences [Resnick and Varian, 1997]. Ils reposent pour cela sur une variété de techniques permettant de suggérer des items susceptibles d'être utiles ou d'intéresser un utilisateur [Ricci *et al.*, 2021]. Le terme *item* est communément employé pour désigner les ressources que le système doit recommander. La nature de ces items varie en fonction du contexte d'application. Des produits pour le commerce en ligne, aux films ou séries pour la vidéo à la demande, en passant par les musiques sur les plateformes d'écoute musicale, ou encore les news dans le secteur de l'actualité [Ko *et al.*, 2022].

Quel que soit le contexte d'application, ces systèmes de recommandation reposent sur la modélisation des préférences des utilisateurs afin d'identifier les items les plus pertinents pour chacun d'entre eux [Ricci *et al.*, 2021]. Considérant un ensemble d'utilisateurs U, et un ensemble d'items I, l'objectif est donc de trouver l'item $i \in I$ (ou l'ensemble des items) correspondant aux besoins d'un utilisateur $u \in U$. Cet objectif peut être formulé selon l'Équation (4.1) [Adomavicius and Tuzhilin, 2005] :

$$i^* = \underset{i \in I}{\operatorname{argmax}}\, v(u, i) \tag{4.1}$$

La fonction v, centrale dans la tâche de recommandation, est une fonction d'utilité d'un item i pour l'utilisateur u. Cette fonction permet généralement de calculer un score de pertinence permettant d'évaluer à quel point l'item est proche des préférences de l'utilisateur. Lorsque plusieurs recommandations sont fournies, elle peut être appliquée plusieurs fois afin d'identifier les k items ayant les scores les plus élevés, notés $top - k$.

Pour évaluer au mieux ces préférences, une hypothèse simple est émise par la communauté : les préférences passées d'un utilisateur peuvent être utilisées pour prédire les préférences fu-

tures de ce même utilisateur. Pour comprendre les besoins spécifiques d'un utilisateur u, une étape essentielle de modélisation utilisateur est appliquée. Cette modélisation repose sur l'exploitation de données potentiellement multiples [He *et al.*, 2023], dont deux types se distinguent [Jawaheer *et al.*, 2014] :

— Les **données explicites** qui correspondent aux retours explicites des utilisateurs par l'intermédiaire d'actions qu'ils effectuent sur le service (*likes*, notes, commentaires, *etc.*).

— Les **données implicites** qui correspondent aux traces d'interaction collectées au cours du temps pour chaque utilisateur, comme les clics, l'historique des pages consultées, le temps passé sur chaque page, *etc.*

Les données collectées sont parfois redondantes, imprécises, incomplètes, *etc.* Une fois collectées elles sont donc nettoyées, transformées et pré-traitées avant d'être fournies en entrée des systèmes de recommandation. Lors de l'exploitation de ces données, les systèmes font face à un paradoxe : la quantité de données globale est massive, mais la quantité de données propre à chaque utilisateur est relativement réduite [Ricci *et al.*, 2011]. La façon dont ces données sont exploitées pour identifier les items à recommander à chaque utilisateur, et tenant compte de ce paradoxe, dépend ainsi des approches appliquées. Les principales approches sont détaillées dans la section suivante. Cependant, j'ai fait le choix de fournir une description globale de ces approches, sans entrer dans les détails, car la contribution présentée dans ce chapitre ne consiste pas en une approche de recommandation.

4.1.2 Approches

Historiquement, les approches de recommandation sont classées en deux familles en fonction de leur mode de fonctionnement [Balabanović and Shoham, 1997] : les algorithmes de filtrage collaboratif [Resnick *et al.*, 1994, Ekstrand *et al.*, 2011] et les algorithmes de filtrage par contenu [Pazzani and Billsus, 2007, Lops *et al.*, 2011]. La combinaison des deux approches donne naissance aux modèles hybrides [Burke, 2002] (Figure 4.1).

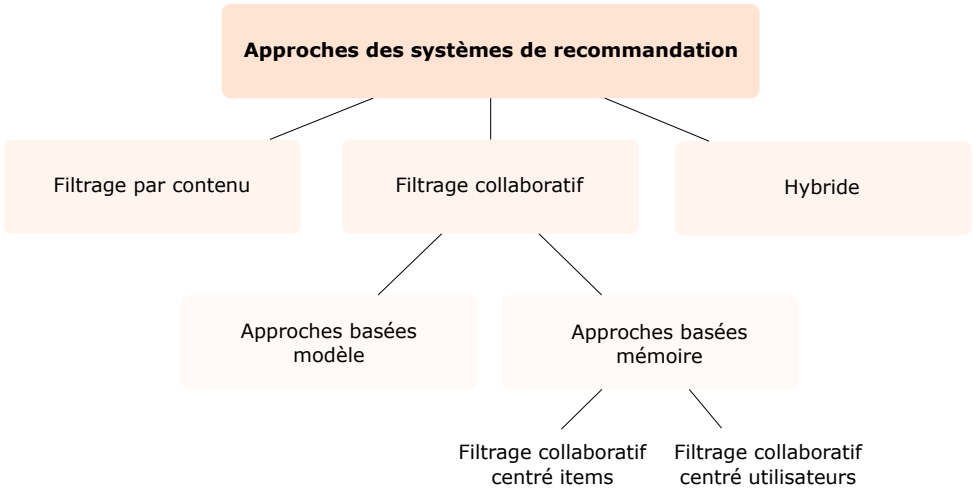


FIGURE 4.1 – Catégorisation des approches de recommandation

Filtrage collaboratif

Le filtrage collaboratif a été décrit dans les années 1990 [Resnick *et al.*, 1994], et est l'approche pionnière de la recherche sur les systèmes de recommandation. Il repose sur l'exploitation des préférences d'un utilisateur cible, implicites ou explicites, et sa comparaison avec celles d'autres utilisateurs du système. Les préférences sont généralement évaluées au travers des notes attribuées par les utilisateurs aux items. Pour prédire la pertinence d'un item i pour un utilisateur u_1, l'hypothèse suivante est posée : si les utilisateurs u_1 et u_2 ont ou ont eu des intérêts similaires sur certains items, ils auront probablement des intérêts similaires sur d'autres items (Figure 4.2).

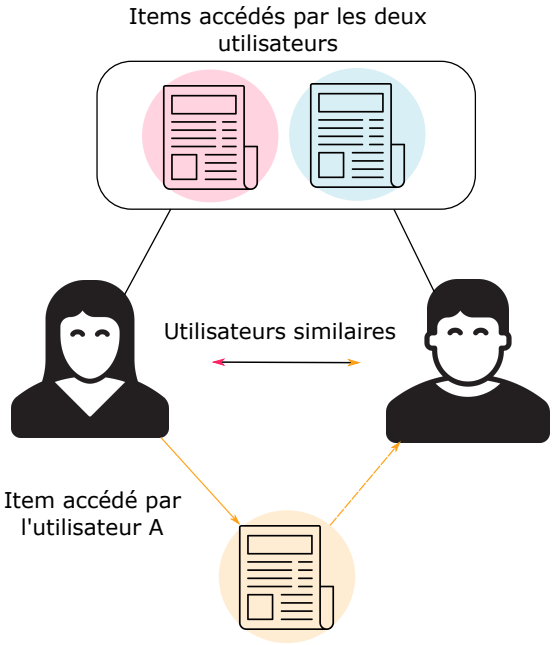


FIGURE 4.2 – Principe du filtrage collaboratif

Parmi les algorithmes de filtrage collaboratif, les ***approches basées mémoire*** sont distinguées des ***approches basées modèle***. Les approches basées mémoire reposent sur l'utilisation de l'ensemble des données sur les préférences des utilisateurs. Pour cela, les utilisateurs (***filtrage collaboratif centré utilisateurs***) ou les items (***filtrage collaboratif centré items***) qui sont les plus similaires à l'utilisateur cible sont identifiés. Pour quantifier cette similarité, des mesures telles que la similarité *cosinus*, le coefficient de corrélation de Pearson, ou encore des approches reposant sur l'identification des plus proches voisins (K-Nearest Neighbors) peuvent être appliquées [Fkih, 2022]. Une fois les voisins identifiés (utilisateurs ou items), ils sont exploités de façon à prédire les préférences de l'utilisateur cible [Ko *et al.*, 2022]. Cependant, utilisant une large quantité de données, les approches basées mémoire sont très coûteuses computationnellement. Ainsi, les approches basées modèles tentent de répondre à cette limite et à réduire les coûts de calcul en exploitant des approches d'apprentissage automatique et de fouille de données pour identifier les items les plus proches des préférences d'un utilisateur. Le modèle est ainsi entraîné à partir des préférences passées de l'utilisateur, et ne nécessite pas l'intégralité des données pour fonctionner. Les principales approches appliquées sont les règles d'association, le clustering, la régression, les classifieurs bayésiens, les arbres de décisions et la factorisation de matrices [Patel *et al.*, 2017]. Dernièrement, les approches d'apprentissage profond (*deep learning*) se sont largement développées dans la littérature et sont notamment appliquées à la recommandation

[Batmaz *et al.*, 2019]. Leurs performances et supériorité, pourtant largement mises en avant dans la littérature, restent cependant à confirmer [Jannach *et al.*, 2020]. [Ferrari Dacrema *et al.*, 2019] ont par exemple tenté de reproduire 18 approches de recommandation reposant sur des approches neurales. Seules 7 ont pu être reproduites, et pour 6 d'entre elles les performances ont été dépassées par des approches de recommandation plus simples telles que l'approche des plus proches voisins. Les modèles profonds peuvent donc présenter une limite liée à leur reproductibilité et être dépassés par des modèles reposant sur des approches plus frugales.

Ainsi, **le filtrage collaboratif permet de fournir des recommandations sans connaître ou analyser précisément les items recommandés**. C'est l'usage et les interactions avec ces items qui est mis en avant et privilégié, et non pas les items eux-mêmes. Cette spécificité permet d'adapter les approches de filtrage collaboratif dans de nombreux domaines, y compris certains pour lesquels les connaissances sont limitées. Bien qu'ayant été introduites depuis plusieurs années, ces approches restent aujourd'hui toujours largement répandues dans la littérature [Ricci *et al.*, 2021], et continuent d'être développées avec l'avènement des approches profondes.

Le filtrage collaboratif présente néanmoins un certain nombre de limites, dont une liste non exhaustive est donnée ci-après :

— **Démarrage à froid *(Cold Start)*** [Lika *et al.*, 2014] : ce phénomène apparaît lorsque de nouveaux utilisateurs utilisent un système et que les informations les concernant sont très peu nombreuses. Il devient alors difficile de générer des recommandations pertinentes. La même problématique s'applique lorsqu'un item est introduit dans le système, avec lequel aucun utilisateur n'a interagi. Il devient alors difficile de le recommander.

— **Passage à l'échelle *(Scalability)*** [Papagelis *et al.*, 2005] : les approches basées mémoire sont très coûteuses en calcul car elles évaluent la similarité entre toutes les paires d'items et d'utilisateurs. Une mise en application dans un contexte réel nécessitant de traiter une large quantité de données peut ainsi rapidement entraîner des temps de calcul considérables.

— **Manque de données *(Sparsity)*** [Sarwar *et al.*, 2000] : certains systèmes ont un catalogue d'items très étendu, dont un grand nombre sont très peu évalués. Il devient alors difficile de fournir des prédictions pour ces items.

— **Utilisateurs *Grey Sheep*** [Claypool *et al.*, 1999] : certains utilisateurs ont des préférences très éloignées de celles des autres utilisateurs, et il devient difficile de leur fournir des recommandations adaptées.

— **Items *Long Tail*** [Park and Tuzhilin, 2008] : certains items peu populaires, et peu consultés par les utilisateurs, ne sont jamais recommandés.

Filtrage par contenu

À l'inverse des approches de filtrage collaboratif, les systèmes de recommandation reposant sur le filtrage par contenu exploitent uniquement les caractéristiques des items disponibles à recommander. Ces derniers sont analysés, et comparés aux items précédemment accédés par l'utilisateur. L'objectif est ainsi d'identifier les items qui se rapprochent au plus près des intérêts de l'utilisateur cible [Pazzani and Billsus, 2007], sans tenir compte des autres utilisateurs (Figure 4.3).

Le développement de ces systèmes de recommandation repose sur une caractérisation du contenu disponible. Pour cela, les approches exploitent le contenu lui-même ou les métadonnées qui fournissent une variété d'informations. La nature de ces contenus dépend de celle de l'item :

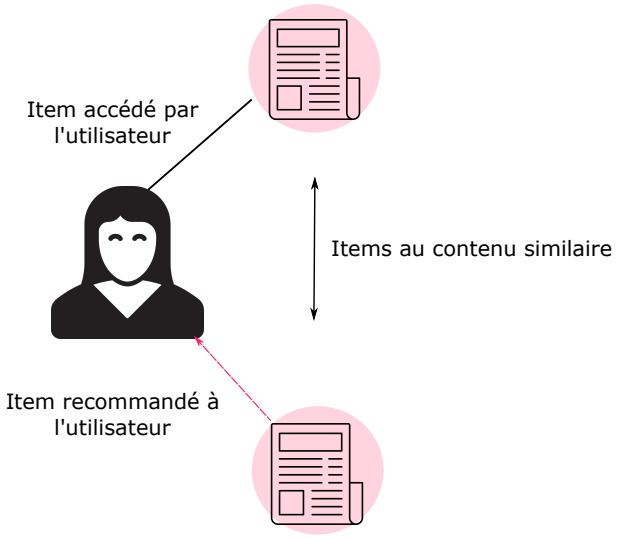


FIGURE 4.3 – Principe du filtrage par contenu

un film pourra être identifié par son titre, ses acteurs, son réalisateur, tandis qu'une musique sera plutôt identifiée par son genre musical, et son artiste. Ces informations peuvent être de formes multiples (valeurs binaires, numériques, catégorielles, textuelles, *etc.*) et requièrent des méthodes d'analyse adaptées. La représentation vectorielle des données est souvent privilégiée lorsqu'elle est possible, mais les données textuelles nécessitent des traitements souvent plus complexes. Des approches de traitement automatique des langues peuvent être adoptées pour donner une représentation numérique au texte [Lops *et al.*, 2011]. Ensemble, toutes les données exploitées et analysées permettent de caractériser les items, puis d'identifier les plus similaires aux préférences des utilisateurs.

Un système de recommandation basé sur le filtrage par contenu repose ainsi sur 3 étapes principales et indépendantes [Lops *et al.*, 2011]. La première étape consiste à **analyser le contenu** en pré-traitant les informations disponibles sur les items afin d'en donner une représentation exploitable par le système de recommandation. Cette représentation est fournie en entrée des étapes suivantes. La seconde étape correspond à la **création du profil de l'utilisateur cible** à partir de la représentation des items avec lesquels il a interagi. La troisième et dernière étape consiste alors à **filtrer les items disponibles** afin d'identifier ceux à recommander. Pour cela, le profil utilisateur est comparé à la représentation des items candidats (*c.-à-d.* qui peuvent être recommandés à l'utilisateur cible), afin d'identifier ceux qui sont les plus similaires aux intérêts de l'utilisateur, et donc les plus susceptibles de correspondre à ses attentes.

Ainsi, **l'approche de filtrage par contenu, comme le filtrage collaboratif, exploite les données d'interaction, mais se différencie par la représentation des items disponibles, qui sont analysés avant d'être recommandés.** Cette particularité permet de répondre à certaines des limites énoncées du filtrage collaboratif (Section 4.1.2) [Thorat *et al.*, 2015]. En particulier, elle permet de répondre au problème de démarrage à froid (*cold-start*) pour les items : lorsqu'un nouvel item est recommandé, ce dernier peut être représenté puis recommandé, sans interactions préalables de la part des utilisateurs. Cependant, le problème de démarrage à froid pour les utilisateurs persiste : lorsqu'un nouvel utilisateur utilise le système, son profil utilisateur est construit sur la base de peu d'interactions, ce qui limite la qualité des recommandations. Cependant, le profil utilisateur étant construit pour chaque utilisateur individuellement, il n'existe

pas d'utilisateurs *grey sheep* à qui le système peine à faire des recommandations pertinentes. Cette approche de recommandation individualisée permet également de répondre à la limite de manque de données (*sparsity)*, puisqu'elle est applicable même s'il n'y a que très peu d'utilisateurs utilisant le système et fournissant des données. Finalement, en ne comparant que les utilisateurs aux items, et pas les utilisateurs entre eux, l'approche par contenu permet de réduire les coûts de calcul, ce qui peut simplifier le passage à l'échelle (*scalability*).

Bien que répondant à certaines des limites du filtrage collaboratif, le filtrage par contenu présente également des limites qui lui sont propres :

— **Dépendance à la qualité des données** [Heinrich *et al.*, 2021] : la qualité des recommandations dépend de la qualité des données disponibles à propos des contenus à recommander. En effet, des données de mauvaise qualité, partielles ou biaisées ne permettent pas de caractériser correctement les ressources, ni de construire un modèle utilisateur pertinent, ce qui limite la qualité des recommandations.

— **Limite de l'analyse des items** [Lops *et al.*, 2011] : la représentation des contenus peut parfois être incomplète ou inadaptée, impactant la qualité des recommandations. Pour les données textuelles, les contenus multilingues ou l'ambiguïté de certains termes peuvent par exemple être difficilement représentés.

— **Sur-spécialisation *(overspecialization)*** [McNee *et al.*, 2006] : la recommandation des items les proches des préférences de l'utilisateur tend à réduire la diversité des recommandations fournies.

Approches hybrides

Les approches de filtrage collaboratif et de filtrage par contenu, bien que différentes, sont considérées comme complémentaires [Adomavicius and Tuzhilin, 2005]. Ainsi, pour tirer profit des deux approches, et tenter de répondre à leurs limites respectives, des approches hybrides ont émergé au début des années 2000 [Burke, 2002]. Au total, ce sont sept approches d'hybridation principales qui ont été caractérisées dans la littérature [Burke, 2002, Roy and Dutta, 2022]. Ces dernières sont détaillées dans le Tableau 4.1.

Ces approches hybrides permettent de répondre aux différentes limites précédemment citées des approches classiques de filtrage collaboratif et de filtrage par contenu (démarrage à froid, surspécialisation, manque de données, *etc.*). Elles représentent désormais la majorité des approches de recommandation proposées dans la littérature [Çano and Morisio, 2017].

Néanmoins, bien que certains modèles de recommandation permettent de répondre aux différentes limites des approches présentées, il subsiste des limites relatives à la qualité des recommandations fournies, qui sont discutées à la suite de cet état de l'art.

4.1.3 Évaluation des systèmes de recommandation

La qualité des systèmes de recommandation est évaluée au travers de cadres d'évaluation reposant sur le calcul de métriques de performances. Le choix de ces métriques dépend notamment du type d'expérimentation menée, qui conditionne la nature et la quantité de données collectées [Beel and Langer, 2015]. Trois types d'expérimentations sont décrits dans la littérature.

Les **expérimentations hors-ligne** (*offline*) reposent sur une simulation des comportements utilisateur à partir d'interactions passées contenues dans des jeux de données existants. Ce type d'évaluation est particulièrement adapté à l'évaluation de la qualité algorithmique d'un système, et peut être mis en place rapidement car ne nécessite pas de recruter des utilisateurs. Cependant, les données historiques sur lesquelles se base l'évaluation peuvent être biaisées, ce qui peut

Méthode	**Description**
Meta-niveau (*meta-level*)	Modèle pré-entraîné est donné en entrée d'un second modèle
Combinaison (*feature combination*)	Utilisation conjointe des représentations de différentes approches
Augmentation (*feature augmentation*)	Représentations d'un modèle données en entrée d'un second modèle
Mixage (*mixed*)	Mélange des résultats de différents modèles
Cascade (*cascading*)	Affinage des résultats d'un modèle à l'aide des résultats d'un autre modèle
Échange (*switching*)	Modification du modèle en fonction du contexte courant
Pondération (*weighted*)	Agrégation et pondération des résultats de différents modèles

TABLE 4.1 – Méthodes de filtrage hybride [Burke, 2002]

impacter les résultats de l'évaluation. Par ailleurs, aucun retour d'expérience d'utilisateurs réels n'est fourni, ce qui ne permet pas de garantir la qualité du système de recommandation s'il est amené à être déployé et utilisé dans des conditions réelles.

Les **expérimentations en ligne** (*online*) permettent de répondre à cette limite. Pour ce second type d'expérimentations, le système évalué est en effet utilisé par des utilisateurs réels. Ce type d'évaluation est donc plus adapté à l'évaluation des performances d'un système dans des conditions réelles, et fournit donc de meilleurs indicateurs quant à la qualité du système s'il est amené à être mis en service. Cependant, ce type d'évaluation peut être difficile à mettre en place car nécessite une infrastructure technique adaptée.

Le troisième type d'expérimentation consiste en la réalisation d'**études utilisateurs**, durant lesquelles le système est utilisé par des utilisateurs réels dans des conditions réelles, mais contrôlées. Certaines tâches spécifiques peuvent donc être définies à l'avance et doivent être exécutées par l'ensemble des utilisateurs participant à l'étude. En plus d'éviter de potentiels biais, ces études permettent d'observer les interactions en temps réel, et d'échanger directement avec les utilisateurs, ce qui peut permettre une identification plus rapide des potentielles améliorations à apporter.

Les métriques d'évaluation des systèmes de recommandation se distinguent également par la dimension qu'elles permettent d'évaluer [Zangerle and Bauer, 2022]. Les systèmes de recommandation étaient initialement développés de façon à optimiser l'exactitude, et seule cette exactitude était évaluée (*c.-à-d.* la capacité à correctement identifier les items d'intérêt pour les utilisateurs) [Konstan *et al.*, 1998, Bellogin *et al.*, 2011]. Cependant, l'évaluation unique de cette exactitude s'est montrée insuffisante pour évaluer la qualité d'un système de recommandation [McNee *et al.*, 2006]. De nombreuses autres métriques permettant d'évaluer des dimensions allant au-delà de l'exactitude ont ainsi été définies.

Dans les sections suivantes, je présente un ensemble de métriques permettant d'évaluer ces dimensions diverses. Du fait qu'elles soient nombreuses, je choisis d'en fournir une présentation

non exhaustive, en détaillant uniquement la formule des métriques les plus largement appliquées dans la littérature pour chacune des dimensions.

Métriques d'exactitude

De nombreuses mesures d'exactitude ont été conçues pour l'évaluation hors ligne de la qualité des modèles de recommandation [Aggarwal, 2016]. Suivant le protocole traditionnel d'évaluation de l'apprentissage machine, le jeu de données est ainsi divisé en un ensemble d'entraînement (*train set*), permettant d'entraîner le modèle, et un ensemble de test (*test set*), permettant de l'évaluer en comparant les sorties de l'algorithme de recommandation entraîné aux données réelles. Les différentes métriques d'exactitude existantes sont **basées sur les erreurs**, sur la **précision** ou sur l'**ordre des recommandations** [Ricci *et al.*, 2021].

Métriques basées sur les erreurs. Elles sont souvent préférées lorsqu'il s'agit d'évaluer des recommandations uniques. Les plus répandues sont l'erreur quadratique moyenne, ou *Mean Squared Error* (MSE), et la racine de cette erreur quadratique moyenne, ou *Root Mean Squared Error* (RMSE) [Chicco *et al.*, 2021]. L'erreur quadratique correspond à la valeur, au carré, de l'erreur entre une valeur prédite par le modèle $\hat{y}_i$ et la valeur observée y_i. La MSE est donc la moyenne de toutes ces erreurs pour un ensemble n de mesures (Équation (4.2)).

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2 \tag{4.2}$$

Ces métriques sont notamment utilisées lorsque les notes attribuées par les utilisateurs à différents items sont disponibles, et permettent ainsi d'évaluer si les prédictions faites par le système se rapprochent des notes réellement attribuées par les utilisateurs [Parra and Sahebi, 2013]. Ces métriques sont donc exclusivement applicables lorsque des notes sont attribuées aux items à recommander, ce qui n'est pas le cas dans tous les contextes de recommandation.

Métriques de précision. Elles permettent d'évaluer si les items recommandés à un utilisateur correspondent effectivement à ceux qui l'intéressent. Les items peuvent ainsi être des vrais positifs (VP) (recommandés et pertinents pour l'utilisateur), des vrais négatifs (VN) (non recommandés et non pertinents pour l'utilisateur), des faux positifs (FP) (recommandés mais non pertinents pour l'utilisateur) ou des faux négatifs (FN) (non recommandé mais pertinents pour l'utilisateur). En d'autres termes, cela permet d'évaluer le taux d'items correctement identifiés par le système [Zangerle and Bauer, 2022]. Les métriques de précision sont ainsi plus générales, et applicables dans des contextes variés, notamment lorsque les notes ne sont pas disponibles. Parmi ces métriques, les plus utilisées sont la précision (Équation (4.3)), le rappel (Équation (4.4)), et la combinaison de ces deux métriques en une métrique appelée F1-score (Équation (4.5)) [Gunawardana *et al.*, 2012]. La précision indique à quel point les recommandations fournies sont pertinentes pour l'utilisateur, tandis que le rappel indique à quel point le système couvre les préférences de l'utilisateur. La combinaison des deux métriques dans le F1-score permet ainsi de fournir une évaluation plus globale.

$$Précision = \frac{VP}{VP + FP} \tag{4.3}$$

$$Rappel = \frac{VP}{VP + FN} \tag{4.4}$$

$$F1 = 2 \times \frac{Précison \times Rappel}{Précision + Rappel} \tag{4.5}$$

Ces métriques peuvent être calculées sur l'ensemble des recommandations calculées, ou sur un top-k de recommandations. Elles sont alors notées *Précision@k*, *Rappel@k* et *F1@k* et ne considèrent que les k recommandations ayant les scores de pertinence les plus élevés. De façon alternative, la précision d'un système peut également être représentée à l'aide de la fonction d'efficacité du récepteur, ou *Receiver Operating Characteristic* (ROC curve), qui permet de représenter l'évolution des taux de vrais positifs et faux positifs en fonction du seuil. Le calcul de l'aire sous cette courbe, ou *Area Under the Curve* (AUC), permet d'évaluer la performance globale du modèle. L'ensemble de ces métriques de précision repose sur l'évaluation d'un ensemble de recommandations, mais ne tient pas compte de leur ordre, qu'il peut être important de prendre en compte pour l'évaluation de l'exactitude [Chen *et al.*, 2023].

Métriques d'ordonnancement. Elles sont adaptées aux systèmes de recommandation qui reposent sur des listes ordonnées ou des séquences [Parra and Sahebi, 2013]. Leur objectif est de déterminer si le classement des recommandations est approprié, *c.-à-d.* si les recommandations les plus pertinentes pour l'utilisateur sont celles recommandées en premier. Parmi ces métriques, la plus communément appliquée est la métrique nDCG (*Normalized Discounted Cumulative Gain*) [Järvelin and Kekäläinen, 2002]. Elle compare l'ordonnancement des recommandations par rapport à l'ordonnancement idéal, où tous les items les plus pertinents pour l'utilisateur apparaissent en premier dans la liste des recommandations. Elle fournit ainsi un score de précision avec une pondération décroissante en fonction du rang de l'item dans l'ensemble des recommandations. Ce score est calculé en comparant le DCG (*Discounted Cumulative Gain*), correspondant au gain dans les recommandations, au gain idéal IDCG (*Ideal Discounted Cumulative Gain*). Ce gain est estimé pour chaque item i et est noté rel_i, avec $rel_i = 1$ si l'item i est pertinent pour l'utilisateur (*c.-à-d.* qui est recommandé et accédé par l'utilisateur cible u), et $rel_i = 0$ sinon. Le gain d'une liste de recommandation correspond alors à un vecteur de valeurs booléennes rel, tel que $|rel| = k$, avec k le nombre de recommandations. Le DCG et le IDCG sont calculés à partir de ce vecteur rel selon les Équations (4.6) et (4.7) respectivement, ce qui permet ensuite de calculer la métrique nDCG selon l'Équation (4.8).

$$DCG(rel) = \sum_{i} \frac{2^{rel_i} - 1}{log_2(i+1)} \tag{4.6}$$

$$IDCG(rel) = \sum_{i=1}^{sum(rel)} \frac{1}{log_2(i+1)} \tag{4.7}$$

$$nDCG(rel) = \frac{DCG(rel)}{IDCG(rel)} \tag{4.8}$$

D'autres métriques permettent d'évaluer l'exactitude des recommandations en tenant compte de leur rang, notamment la métrique *Mean Reciprocal Rank* (MRR) dont le score indique à quel point les items pertinents pour un utilisateur sont recommandés dans les rangs les plus hauts.

Comme je l'ai indiqué au début de cette section, le choix des métriques se fait en fonction du type d'évaluation souhaité. Les métriques d'exactitude présentées jusqu'ici sont adaptées à une évaluation hors ligne, et d'autres métriques sont adaptées à l'évaluation dans le cadre

d'études utilisateur ou d'expérimentations en ligne. Le travail que je présente dans ce manuscrit reposant sur une étude hors ligne, je ne donne qu'un bref aperçu des métriques les plus utilisées lors d'évaluations en ligne. La mesure Click-Through Rate (CTR) (Équation (4.9)) permet par exemple de mesurer le taux de clics sur les recommandations lorsque celles-ci sont fournies dans des conditions réelles d'utilisation.

$$CTR = \frac{Nombre\ de\ clics}{Nombre\ total\ d'impressions} \times 100 \tag{4.9}$$

Le nombre de clics correspond au nombre total de clics sur les recommandations, tandis que le nombre total d'impressions correspond au nombre de fois où les recommandations ont été affichées. Ainsi, plus le taux est élevé, plus les recommandations répondent aux attentes des utilisateurs. De façon similaire, le taux de conversion ou taux d'adoption mesure le nombre de clics qui conduisent effectivement à la consommation d'items recommandés [Zangerle and Bauer, 2022]. Dans un contexte économique, ces métriques sont également utiles pour évaluer la valeur commerciale d'un système de recommandation [Jannach and Jugovac, 2019].

Les mesures d'exactitude présentées dans cette section ont fait leurs preuves pour évaluer la pertinence des recommandations, mais elles souffrent de limitations. En particulier, la précision s'est avérée insuffisante pour évaluer et expliquer la satisfaction des utilisateurs [Zhou *et al.*, 2010, Maksai *et al.*, 2015]. [Said *et al.*, 2013] ont par exemple montré que deux algorithmes de recommandation aux performances d'exactitude totalement différentes lors d'une évaluation hors ligne, peuvent être évalués comme ayant une utilité identique lors d'une évaluation en ligne. En ce sens, [Jannach and Adomavicius, 2016] expliquent que les systèmes de recommandation répondent à des objectifs particuliers et propres à chaque utilisateur. **Les recommandations qui sont les plus exactes ne sont donc pas nécessairement celles qui sont les plus utiles pour les utilisateurs [McNee *et al.*, 2006]. L'évaluation des systèmes de recommandation doit donc reposer sur des dimensions allant au-delà de l'exactitude.**

Métriques allant au-delà de l'exactitude

Pour répondre à l'évaluation partielle permise par les métriques d'exactitude, les cadres d'évaluation ont récemment été enrichis de métriques complémentaires, comme la **diversité** [Kunaver and Požrl, 2017], la **nouveauté** [Zhang, 2013], la **sérendipité** [Ge *et al.*, 2010], et l'**équité** [Wang *et al.*, 2023]. Elles permettent d'évaluer des aspects essentiels des systèmes de recommandation, notamment des aspects liés à l'humain. En plus d'être proches des préférences, les recommandations doivent en effet permettre de répondre à d'autres besoins des utilisateurs comme la découverte ou l'enrichissement.

Diversité. Elle a été introduite pour la première fois par [Smyth and McClave, 2001], qui ont mis en évidence le manque de diversité dans les recommandations produites par les approches centrées sur l'exactitude. La diversité est alors définie comme la dissimilarité (1-similarité) des items recommandés. Cette similarité peut-être évaluée avec des mesures de similarité comme la similarité *cosinus*. Reposant sur cette mesure de dissimilarité, la métrique de diversité intra-liste (ILD) [Smyth and McClave, 2001, Ziegler *et al.*, 2005], permet d'évaluer la diversité d'une liste de recommandations R (Équation (4.10)).

$$ILD(R) = \frac{1}{|R|(|R|-1)} \sum_i \sum_j (1 - Similarity(i,j)) \tag{4.10}$$

Cette mesure ILD est la plus communément appliquée dans la littérature, mais d'autres métriques viennent la compléter. La diversité relative, ou *Relative Diversity* (RD), permet par exemple d'évaluer la diversité apportée par un item spécifique au sein d'un ensemble de recommandations [Bradley and Smyth, 2001].

Une mesure de diversité sensible à l'ordre des recommandations, appelée *Expected Intra-List Diversity* (EILD) [Vargas and Castells, 2011] permet d'évaluer la diversité attendue d'une liste de recommandations en tenant compte du rang de chaque item dans le calcul de diversité. [Clarke *et al.*, 2008] ont également proposé une métrique inspirée de nDCG, appelée α-nDCG, tenant compte de la redondance et de différents aspects des items d'une liste de recommandation, en faisant une métrique de diversité. Ces aspects correspondent à des dimensions permettant de caractériser les items, et sont donc dépendants du contexte d'application (genres de musiques, thématiques de films, *etc.*). Reposant sur cette notion d'aspects, [Zhai *et al.*, 2015] ont défini une métrique appelée *Subtopic Recall* (S-Recall), définie dans l'Équation (4.11). Cette métrique permet d'évaluer la proportion des aspects A présents dans le $top-k$ des recommandations.

$$S - Recall@k = \frac{|\cup_i aspect(i)|}{|A|} \tag{4.11}$$

Finalement, la *diversité agrégée* permet d'évaluer la diversité des recommandations fournies à l'ensemble des utilisateurs [Adomavicius and Kwon, 2011].

Nouveauté. Proche de la notion de diversité, elle a été rapidement discutée et identifiée comme importante dans un contexte de filtrage de l'information [Baeza-Yates *et al.*, 1999]. Elle correspond à la présence d'items nouveaux dans les recommandations [Ricci *et al.*, 2011]. Bien que le concept paraisse simple, les définitions associées sont multiples. En particulier, la nouveauté peut être étudiée du point de vue des items, ou bien de celui des utilisateurs. Lorsque considérée du point de vue des items, la nouveauté correspond notamment à la recommandation d'items peu populaires, contenus dans la *long tail*. La nouveauté est alors abordée selon la popularité des items, où la popularité d'un item, notée $p(i)$, correspond à la fraction d'utilisateurs ayant noté ou accédé à l'item parmi l'ensemble des utilisateurs du système [Kaminskas and Bridge, 2016]. La nouveauté d'une liste de recommandation R est alors évaluée selon l'Équation (4.12).

$$Novelty(R) = \frac{\sum_i -log_2 p(i)}{|R|} \tag{4.12}$$

Par ailleurs, lorsque la nouveauté est considérée du point de vue des utilisateurs, un item est évalué comme étant nouveau s'il est inconnu, pertinent, et dissimilaire aux items avec lesquels l'utilisateur a déjà interagi [Vargas, 2015].

Lorsqu'en plus d'être nouveaux, les items sont inattendus pour l'utilisateur, la notion de sérendipité est généralement évoquée[Ricci *et al.*, 2011]. Plus précisément, des recommandations présentant de la sérendipité contiennent des items surprenants mais satisfaisants pour l'utilisateur [Ge *et al.*, 2010]. L'évaluation de la sérendipité repose ainsi sur une mesure de l'inattendu (*unexpectedness*) [Herlocker *et al.*, 2004], et peut donc être évaluée en fonction de la fraction d'items inattendus mais pertinents pour l'utilisateur au sein d'une liste de recommandations [Kaminskas and Bridge, 2016].

Équité. L'équité est associée à de nombreuses définitions, et concerne à la fois les utilisateurs et les items. L'équité à l'égard des utilisateurs est définie comme l'absence de biais, de préjudice, de favoritisme ou de mauvais traitement à l'égard d'individus, de groupes, classes ou catégories sociales à partir des caractéristiques qui leurs sont propres [Elahi *et al.*, 2022]. Suivant cette définition, la métrique de disparité des recommandations (*Recommendation Disparity*) [Leonhardt *et al.*, 2018] permet d'évaluer si la qualité des recommandations est la même pour l'ensemble des utilisateurs. Une autre métrique d'équité ayant gagné l'intérêt des chercheurs ces dernières années est la métrique de calibration [Steck, 2018], abordant l'équité vis-à-vis des intérêts des utilisateurs, ainsi que leurs proportions. La calibration permet en effet d'évaluer dans quelle mesure la liste de recommandations est compatible avec les intérêts de l'utilisateur. Ces intérêts sont définis sur un ensemble d'aspects caractérisant les items. La mesure de calibration compare ainsi la distribution de probabilité p des recommandations sur ces aspects, à la distribution de probabilité q des intérêts de l'utilisateur sur ces mêmes aspects (Équation 4.13). Ainsi, plus la valeur de calibration est proche de 0, plus les distributions de probabilité sont proches et donc plus la liste de recommandation est équitable.

$$C_H(p, q) = \frac{1}{2} \times ||\sqrt{p} - \sqrt{q}||_2 \tag{4.13}$$

Lorsque l'équité est évaluée à l'égard des items, elle est définie comme un traitement impartial de l'ensemble de ces items, qui ont tous la même opportunité d'être recommandés, indépendamment de leurs caractéristiques [Li *et al.*, 2023]. Des mesures de couverture permettent d'évaluer dans quelle mesure un système est capable de recommander l'ensemble des items disponibles [Ge *et al.*, 2010].

Pour résumer, les métriques d'évaluation de l'état de l'art sont nombreuses, et permettent d'étudier une large variété de concepts. Lors de la conception d'un système, le choix des métriques dépend à la fois des données utilisées, du type d'évaluation mis en place et des objectifs de recherche [Zangerle and Bauer, 2022]. Cependant, dans tous les cas **il est primordial de proposer une évaluation exhaustive du système reposant sur un cadre d'évaluation multi-métrique [Jannach and Adomavicius, 2016]**. Une telle évaluation permet d'assurer qu'en plus de fournir des recommandations répondant aux préférences des utilisateurs, le système de recommandation offre une expérience utilisateur optimale.

4.2 Systèmes de recommandation de news : spécificités et challenges

4.2.1 Approches

L'évolution vers le partage et la consommation de l'actualité en ligne a permis aux lecteurs d'accéder à un grand nombre d'informations chaque jour. C'est dans ce contexte qu'ont été mis au point les systèmes de recommandation de news [Raza and Ding, 2022].

La recommandation de news est une application directe de la tâche de recommandation. À partir de l'historique des consommations d'un utilisateur cible, et de la modélisation associée, l'objectif est donc d'identifier les news les plus proches des préférences des utilisateurs, et qui sont susceptibles de l'intéresser [Wu *et al.*, 2023]. Les systèmes de recommandation de news reposent ainsi sur les mêmes approches de recommandations que celles présentées dans la première section de cet état de l'art, à savoir les approches de filtrage collaboratif, de filtrage par contenu et hybrides.

Cependant, bien que les approches de filtrage collaboratif soient les plus répandues dans la littérature, les approches de filtrage par contenu sont les plus représentées dans le domaine des news [Raza, 2021]. Cette différence s'explique notamment par l'information textuelle contenue dans les news, qui est exploitable par les approches de filtrage par contenu. Ces dernières reposent notamment sur l'application d'approches issues du domaine de traitement automatique des langues, permettant de donner une représentation numérique au contenu textuel. Les approches récentes les plus répandues consistent à donner une représentation vectorielle des mots, phrases ou documents, dans des objets appelés *embeddings*. La représentation des contenus dans l'espace vectoriel résultant permet ainsi de les comparer, et notamment d'évaluer leur proximité sémantique.

Une modélisation traditionnelle, appelée *Term Frenquency-Inverse Document Frequency* (TF-IDF) [Sparck Jones, 1972] et permettant de modéliser le contenu à partir de mots clés, peut par exemple être appliquée dans le domaine des news [Yunanda *et al.*, 2022].

Au-delà d'une modélisation à partir des mots-clés, il est possible d'appliquer une modélisation thématique des news. L'approche de *Latent Dirichlet Allocation* (LDA) [Blei *et al.*, 2003] permet par exemple de représenter chaque news sous la forme d'une distribution thématique. [Luostarinen and Kohonen, 2013] proposent d'appliquer LDA avec différentes approches de filtrage standard (classification naïve bayésienne, plus proches voisins et régression linéaire régularisée). Les résultats montrent une amélioration des performances de recommandation lorsque le contenu des news est représenté avec LDA.

Cette modélisation thématique du contenu peut également être complétée en considérant d'autres dimensions comme la gravité des thématiques, les sentiments exprimés ou encore la complexité linguistique [Tintarev *et al.*, 2018]. Récemment, des représentations reposant sur des concepts plus fins que les mots ou phrases sont appliquées. [Wu *et al.*, 2020a] proposent par exemple d'avoir une modélisation des sentiments exprimés dans les news pour la recommandation. La qualité des recommandations fournies est donc fortement corrélée à la qualité de la représentation des news, ce qui nécessite une collaboration avec les experts du traitement automatique des langues [Dufraisse *et al.*, 2022].

Au-delà de la représentation du contenu, les systèmes de recommandation de news reposant sur des modèles de *deep-learning* se développent [Wu *et al.*, 2023]. Le modèle de recommandation DKN (*Deep Knowledge-aware Network*) [Wang *et al.*, 2018] exploite par exemple des *embeddings* des news construits à l'aide d'un réseau neurones convolutionnel à des connaissances extraites de graphes de connaissances. Les recommandations sont ainsi calculées en combinant à la fois les représentations des news et les informations issues des graphes qui permettent de créer une représentation complexe des news. Par ailleurs, l'approche LSTUR (*Neural Recommendation with Long- and Short-term User Representations*) [An *et al.*, 2019] repose sur l'utilisation de réseaux de neurones pour capturer les préférences à courts et longs termes des utilisateurs et fournir les recommandations. Ces algorithmes reposent tous sur du filtrage par contenu, et exploitent les capacités des réseaux de neurones profonds pour capturer les interactions complexes entre les utilisateurs, les articles et les contextes, afin de fournir des recommandations personnalisées. L'objectif de la thèse présentée dans ce manuscrit n'étant pas de développer de tels systèmes, ils ne seront pas plus détaillés.

Bien que la recommandation de news basée sur le contenu soit plus répandue, d'autres approches sont également présentées dans la littérature. [Alabduljabbar *et al.*, 2023] ont par exemple montré que le couplage entre une approche de filtrage collaboratif ou de filtrage par contenu à des informations contextuelles permettait d'améliorer les performances de la recommandation de news. Ces systèmes de recommandation contextuels (Context-Aware Recommender

Systems) permettent ainsi d'adapter les recommandations en fonction du contexte des utilisateurs afin d'améliorer l'expérience utilisateur [Adomavicius and Tuzhilin, 2010]. Les contextes auxquels peuvent s'adapter les systèmes sont multiples : contexte physique (position géographique, période spécifique de l'année, *etc.*), contexte personnel (genre, âge, environnement social, humeur, *etc.*), ou encore contexte technique (type d'outil utilisé, type de données, *etc.*) [Ferdousi *et al.*, 2017]. Dans le contexte des news, [Viana and Soares, 2016] proposent d'exploiter des informations de géolocalisation pour adapter les recommandations de news. D'autres facteurs comme la fatigue des utilisateurs, correspondant à une perte d'intérêt pour les recommandations, peuvent également être pris en compte et permettre d'adapter les recommandations de news [Ma *et al.*, 2016].

4.2.2 Spécificités liées au contexte des news

Bien que suivant les étapes de fonctionnement classiques d'un système de recommandation, les systèmes de recommandation de news présentent certaines spécificités qui leurs sont propres [Lunardi *et al.*, 2020, Raza and Ding, 2022] :

- **Durée de vie des news :** les news ont une durée de vie très courte, deviennent rapidement obsolètes et sont constamment remplacées par de nouvelles publications. Les systèmes doivent donc être adaptés à cette dynamicité du contexte.
- **Quantité de données :** le nombre de news publiées chaque jour est très important, le catalogue d'items disponibles étant alors très étendu. Les systèmes doivent permettre de traiter une large quantité de données.
- **Temps de consommation :** les news sont généralement courtes et consommées rapidement par les utilisateurs. Les systèmes doivent donc permettre de fournir des recommandations pertinentes dans des temps relativement courts.
- **Qualité du contenu :** la diffusion de fausses informations (*fake news*) complique le processus de recommandation qui doit être en mesure d'évaluer la qualité des items.

Le développement de systèmes de recommandation doit donc tenir compte de ces différentes spécificités et des défis associés.

À cela s'ajoute un autre enjeu, commun à tous les systèmes de recommandation : le respect de l'éthique. En effet, [Milano *et al.*, 2020] soulignent que le développement de systèmes de recommandation ne doit pas seulement être motivé par le rendement ou par un modèle économique, mais aussi par le respect des principes éthiques. Ces systèmes, de part leur utilisation quotidienne et ubiquitaire, ont un impact important sur les utilisateurs. **Dans le cas spécifique des systèmes de recommandation de news, qui jouent un rôle démocratique [Helberger, 2021] et impactent la société au travers du renforcement de la polarisation, le respect des aspects éthiques est d'autant plus fondamental.**

Le filtrage automatisé des informations questionne notamment le respect de la *confidentialité*, puisqu'il repose sur l'analyse de nombreuses données personnelles propres à chaque utilisateur. Ces données potentiellement sensibles traduisent une certaine vulnérabilité des utilisateurs, dont les informations sont collectées, parfois inégalement, et exploitées pour le processus de recommandation [Bonicalzi *et al.*, 2023]. Pour satisfaire le respect de la confidentialité, [Himeur *et al.*, 2022] proposent notamment de définir des directives pour la collecte et le stockage sécurisés des données. Au-delà de la collecte et du stockage, c'est également leur utilisation qui doit être raisonnée, afin d'éviter toute violation de confidentialité.

Pour compléter, la notion de transparence des approches de recommandation est également discutée dans la littérature. En effet, ces systèmes agissent généralement comme des boîtes noires (*black boxes*), et leur fonctionnement n'est pas explicitement détaillé aux utilisateurs

[Elahi *et al.*, 2022]. Les choix effectués par les systèmes, et les mécanismes sous-jacents, doivent cependant pouvoir être expliqués. Ces informations peuvent ainsi permettre aux utilisateurs d'avoir une meilleure compréhension de l'approche de recommandation, leur permettant éventuellement d'adapter leurs interactions de façon à améliorer leur expérience utilisateur. En ce sens, [Diakopoulos and Koliska, 2017] mettent en avant le besoin de lignes directrices permettant d'assurer la transparence algorithmique, qu'ils définissent comme la divulgation d'informations sur les algorithmes pour permettre la surveillance, le contrôle, la critique ou l'intervention des parties prenantes. Ces auteurs parlent d'une transition des boîtes noires (*black boxes)* à des boîtes transparentes (*glass boxes*), qui sont aisément explicables.

Finalement, l'assistance des utilisateurs dans leur recherche d'information questionne également quant à leur *autonomie* : les systèmes ont le pouvoir de remodeler, au travers des recommandations, les décisions de leurs utilisateurs [Bonicalzi *et al.*, 2023]. Ces derniers doivent cependant rester capables de formuler des choix informés. Les systèmes de recommandation ne doivent pas remplacer l'identité des utilisateurs au travers de l'exploitation de leurs préférences, mais davantage contribuer à la construction de cette identité [Floridi, 2011]. Les utilisateurs doivent ainsi rester maîtres de leurs actions, ou choix, et doivent conserver un certain contrôle sur les systèmes [Harambam *et al.*, 2019].

Dans le contexte de la recommandation de news, [Elahi *et al.*, 2022] appellent au développement de systèmes responsables, et notamment au respect de l'*équité* dans les recommandations. **Les systèmes doivent donc veiller à fournir des recommandations de même qualité pour tous les utilisateurs, en veillant à ne pas les influencer.** Finalement, l'une des notions liée à cette idée d'influence, largement explorée dans la littérature et au cœur du sujet de thèse présenté dans ce document, est l'accentuation du phénomène de polarisation résultant du filtrage de l'information. La section 4.3 de cet état de l'art est dédié à cet aspect.

4.2.3 Jeux de données de recommandation de news

La conception d'un système de recommandation de news dépend fortement des données qui sont utilisées : nombre d'utilisateurs, nombre de news, attributs des utilisateurs et des news, activités des utilisateurs, récence, durée, etc.

Pour des raisons de droits d'auteur et de protection de la vie privée, peu de médias en ligne offrent un accès libre au contenu des actualités et à leur consommation par les utilisateurs. Le nombre de jeux de données accessibles au public, même à des fins de recherche, est donc considérablement réduit. En conséquence, de nombreuses études explorent des jeux de données privés qui ne peuvent être ni partagés, ni réutilisés [Karimi *et al.*, 2018].

Dans la plupart des cas, ces jeux de données sont de petite taille et recueillent des informations pour un nombre limité d'utilisateurs, sur de courtes périodes. Pour faire face à cette limite, certains chercheurs effectuent des évaluations sur des jeux de données qui ne sont pas liés aux news, comme le populaire jeu de données MovieLens [Ji *et al.*, 2016]. Cependant, ces jeux de données ne permettent pas de tirer des conclusions solides sur la recommandation de news.

Certains jeux de données dédiés à la recommandation d'actualités sont néanmoins disponibles gratuitement. Un aperçu de ces derniers est donné dans le Tableau 4.2. Il est important de souligner que ces jeux de données contiennent des informations sur les news, mais également sur les interactions des utilisateurs sur ces news [Raza and Ding, 2022]. D'autres domaines, tels que le traitement automatique des langues, sont également intéressés par les jeux de données de news, mais ne contiennent pas d'informations relatives aux interactions utilisateurs et ne sont donc pas exploitables pour la recommandation.

TABLE 4.2 – Description des jeux de données pour la recommandation de news.

Jeu de données	Description	Types de données	Taille	Période
Yahoo Webscope	Plusieurs jeux de données disponibles https://webscope.sandbox.yahoo.com/	Information sur les news Données d'interaction	Dépend du jeu de données	Dépend du jeu de données
Plista [Kille *et al.*, 2013]	Plista & TU Berlin 13 portails de news allemands Accessible à la demande pour la recherche https://www.plista.com/	Informations sur les éditeurs Informations sur les lecteurs Clics et impressions Informations temporelles	17M sessions 70k news 80M impressions 1M clics	30 jours (Juin 2013)
Adressa [Gulla *et al.*, 2017]	Adressavisen et Norwegian U. of Sciences and Technology (NTNU) Versions courte et large https://reclab.idi.ntnu.no/dataset/	Informations sur les news Données d'interactions	*Version large :* 3M utilisateurs, 48k news 27M clics *Version courte :* 15k utilisateurs, 1k news 2.7M clics	*Version large :* 10 semaines (2017) *Version courte :* 1 semaine (2017)
MIND [Wu *et al.*, 2020b]	Microsoft News Website Versions courte et large https://msnews.github.io/	Informations sur les news Logs d'impressions	*Version large :* 1M utilisateurs, 200k news 15M logs *Version courte :* 50k utilisateurs, 60k news 1M logs	*Version large :* 6 semaines (Oct.-Nov. 2019) *Version courte :* 5 semaines (Oct.-Nov. 2019)
Globo.com [de Souza Pereira Moreira *et al.*, 2018]	Portail de news brésilien https://www.globo.com/	Informations sur les news Données d'interactions	3M clics 1.2M sessions 330k utilisateurs, 50k news	2 semaines (Octobre 2017)
Outbrain	Outbrain Click Prediction challenge https://www.kaggle.com/c/outbrain-click-prediction	Informations sur les news Données d'interactions	2 billion visites 700M utilisateurs, 17M clics	2 semaine (Juin 2016)

4.3 Recommandation et polarisation

Les systèmes de recommandation traditionnels transposés dans le domaine de l'actualité fournissent des recommandations guidées par l'exactitude, et ne remettent donc pas en cause les croyances préalables des utilisateurs, en leur proposant des news similaires à celles qu'ils ont consultées précédemment. Comme détaillé dans la première partie de ce manuscrit (Section 1.2.2), il en résulte souvent la création de bulles de filtre [Pariser, 2011], qui sont susceptibles de renforcer le phénomène de polarisation.

Une approche intuitive pour répondre à ce problème consiste à accroître autant que possible la diversité de l'ensemble des news recommandées sans altérer l'exactitude des recommandations [Bernstein *et al.*, 2020]. L'objectif d'une telle approche est de confronter l'utilisateur à un plus large spectre d'opinions [McNee *et al.*, 2006], participant à réduire l'enfermement des utilisateurs au sein de bulles de filtre. Cet apport en diversité peut ainsi également participer à l'adoption d'un débat démocratique sain [Bernstein *et al.*, 2020], pour lequel les utilisateurs agissent comme des citoyens informés au travers de sources d'informations multiples, mettant en avant une diversité de points de vue [Helberger *et al.*, 2018]. Cependant, ces stratégies de diversification visant à rétablir une prise de décision éclairée, et à réduire la polarisation, doivent respecter des recommandations éthiques [Giunchiglia *et al.*, 2021], qui n'influencent pas les utilisateurs et qui sont transparentes, mais aussi équitables [Schelenz, 2021] en respectant les intérêts et préférences des utilisateurs. Ces approches sont détaillées dans les sous-sections suivantes.

4.3.1 Apport en diversité dans les recommandations

L'apport en diversité est un défi commun à tous les domaines de recommandation. Cependant, dans le domaine spécifique de la recommandation de news, cet apport en diversité répond également à des dimensions éthiques et sociétales. C'est le potentiel impact sur la société d'une sur-spécialisation des recommandations qui est abordé. L'apport en diversité permet alors de réduire les phénomènes sociaux tels que la polarisation.

L'apport en diversité fait notamment face au dilemme précision-diversité, selon lequel un gain en diversité s'accompagne inévitablement d'une baisse de la précision [Zhou *et al.*, 2010]. Cependant, privilégier la diversité au détriment de la précision peut entraîner des recommandations moins pertinentes ou moins adaptées aux goûts individuels des utilisateurs. Ainsi, trouver le bon équilibre entre précision et diversité est essentiel pour concevoir des systèmes de recommandation efficaces et satisfaisants pour les utilisateurs. Cela nécessite souvent des approches innovantes qui intègrent à la fois des techniques de recommandation personnalisées et des mécanismes de diversification pour offrir une expérience optimale. Cela a donné naissance à une nouvelle génération de systèmes basés sur la diversité : des systèmes bi-objectifs qui cherchent à optimiser un compromis précision-diversité [Smyth and McClave, 2001]. Ces systèmes partent du principe qu'il existe un équilibre à trouver entre précision et diversité [Di Noia *et al.*, 2014, Ziegler *et al.*, 2005], de sorte que la liste finale de recommandations soit conforme aux préférences antérieures des utilisateurs tout en leur permettant de se confronter à des éléments diversifiés. De nombreux modèles d'apprentissage automatique ont ainsi été développés pour accroître la diversité dans les systèmes de recommandation dans différents domaines d'application tels que le commerce électronique [Hu and Pu, 2011], les services de musique en ligne [L'Huillier *et al.*, 2014, Baracskay *et al.*, 2022], et les réseaux sociaux [Sanz-Cruzado and Castells, 2018].

Approches d'apport en diversité globales

L'apport en diversité est parfois explicitement prise en compte dans la fonction objectif du système [Hurley, 2013, Shi *et al.*, 2012, Su *et al.*, 2013], mais la plupart des études s'appuient sur des stratégies de ré-ordonnancement. Pour limiter les coûts de calcul trop importants de ces stratégies, les heuristiques et approches gloutonnes (*greedy*) sont souvent appliquées [Kaminskas and Bridge, 2016]. Elles s'appuient généralement sur l'approche *Maximal Marginal Relevance* (MMR) [Carbonell and Goldstein, 1998], initialement utilisée dans le domaine de la recherche d'information. La fonction objective de cette approche est présentée dans l'Équation (4.14) :

$$f_{obj}(i, R^*) = (1 - \lambda) \cdot acc(u, i) + \lambda \cdot diversity(i, R) \tag{4.14}$$

où i est un item, u est un utilisateur, R est la liste initiale de recommandation, et R^* est la liste ré-ordonnée. La partie gauche de la fonction traite de la précision, qui peut être évaluée en fonction du score de pertinence calculé par un algorithme de recommandation. La partie droite traite d'une dimension qui va au-delà de la précision, en particulier la diversité, généralement évaluée à l'aide de la *diversité intra-liste* [Ziegler *et al.*, 2005]. Cette fonction objective est donc une combinaison linéaire entre la pertinence et la diversité de la liste de recommandation. Le paramètre λ ($(0 \leq \lambda \leq 1)$) permet de faire varier la pondération de l'importance relative de la précision par rapport à la diversité. Ainsi, plus λ augmente, plus l'importance de la diversité est grande et donc plus les recommandations sont diversifiées. Ainsi, lors de l'application d'une stratégie *greedy* pour apporter de la diversité dans les recommandations, la liste originale R est ré-ordonnée de façon à optimiser cette fonction objectif de l'Équation (4.14). Les éléments optimisant cette fonction sont peu à peu ajoutés à la liste de recommandation finale R^*. Le principe de ce ré-ordonnancement *greedy* permettant d'apporter de la diversité dans les recommandations est détaillé dans l'Algorithme (1).

Algorithme 1 : La stratégie de ré-ordonnancement *greedy*.

Entrée : Liste initiale de recommandations **R** de taille N
Sortie : Liste ré-ordonnée de recommandations **R*** de taille k, $k \leq N$

$R^* = <>$;
tant que $/R^*/ < k$ **faire**
  $i^* = \underset{i \in R \setminus R^*}{argmax} f_{obj}(i, R^*, u)$;
  $R^* = R^* \circ i^*$;
  $R = R \setminus i^*$
fin
retourner R^*;

Cette stratégie de ré-ordonnancement *greedy* ne permet pas d'établir un équilibre entre plus de deux dimensions. Dans un contexte de recommandation multi-objectif, où des dimensions supplémentaires à l'exactitude et à la diversité sont optimisées, l'application de cette approche est donc limitée [Jannach and Abdollahpouri, 2023]. Par ailleurs, bien que permettant un apport conséquent en diversité, **l'approche myopique de ce ré-ordonnancement *greedy* ne permet pas de contrôler la nature de la diversité apportée, mais uniquement sa quantité** [Seymen *et al.*, 2021]. Pour assurer un contrôle plus fin de cet apport, certaines approches personnalisées ont été décrites dans la littérature.

Approches d'apport en diversité personnalisées

Bien que largement discuté, l'impact de l'apport en diversité dans les recommandations de news sur le phénomène de polarisation reste contesté. [Heitz *et al.*, 2022] soulignent que le fait de fournir des recommandations diverses peut entraîner une dépolarisation. À l'inverse, [Bail *et al.*, 2018] expliquent que la diversité peut potentiellement conduire à des points de vue plus extrêmes. Suivant la même intuition, [Wu *et al.*, 2018] affirment que le processus de diversification devrait être personnalisé en fonction des besoins individuels des utilisateurs. Il est alors légitime de s'interroger sur l'impact d'une approche universelle de diversification (*c.-à-d.* un apport en diversité non personnalisé) sur les comportements de consommation de l'information.

Bien qu'elles soient peu appliquées pour la recommandation de news, des approches personnalisées d'apport en diversité ont pourtant été développées. Certaines prennent compte des préférences sous-jacentes des utilisateurs lors de la diversification, elles sont appelées diversification consciente de l'intention (*intent-aware diversification*) [Agrawal *et al.*, 2009, Vargas *et al.*, 2011, Vargas *et al.*, 2012]. Les préférences sont évaluées en fonction des aspects pris en compte dans le processus de diversification. Je rappelle que ces aspects permettent de caractériser les items recommandés. Dans le cadre de la recommandation de news, ils peuvent correspondre par exemple aux sources d'informations, aux catégories de news, *etc.* [Steck, 2018]. L'approche xQuAD (*Explicit Query Aspect Diversification*), initialement développée dans le domaine de la recherche d'information (RI) [Santos *et al.*, 2010], et adaptée aux systèmes de recommandation [Vargas, 2015], permet par exemple d'adapter le calcul du score de pertinence d'un item pour un utilisateur en pondérant par la probabilité que cet item soit associé à un certain aspect.

Certaines études exploitent la notion de sous-profils [Vargas and Castells, 2013], considérant que les utilisateurs peuvent avoir des intérêts différents, représentés par des sous-profils. Dans ces travaux, les recommandations sont formulées en fonction de ces sous-profils, puis combinées pour former la liste finale de recommandations. Plus récemment, [Kaya and Bridge, 2019b] proposent l'approche *Sub-Profile Aware Diversification* (SPAD), qui vise à appliquer une diversification inspirée de l'approche xQuAD, mais en divisant le profil de l'utilisateur en plusieurs sous-profils de façon à apporter davantage de diversité ; [de Campos *et al.*, 2023] présentent l'hybridation des profils utilisateurs à la fois thématiques et temporels pour les recommandations basées sur le contenu ; et [Barkan *et al.*, 2023] s'appuient sur une notion connexe de personas pour augmenter la diversité tout en minimisant la perte de précision. En prenant en compte les sous-profils des utilisateurs, ces travaux personnalisent la diversification. D'autres travaux envisagent également de personnaliser la diversification en fonction de l'inclination des utilisateurs à la diversité : [Eskandanian *et al.*, 2017] identifient par exemple différents groupes d'utilisateurs en fonction de la diversité de leurs interactions à l'aide d'un algorithme de clustering, puis adaptent le niveau de diversité pour les utilisateurs de ces différents groupes. Enfin, certains travaux adaptent l'apport en diversité en fonction de la personnalité [Lu and Tintarev, 2018, Wu *et al.*, 2018], qui est évaluée à l'aide du modèle de personnalité à cinq facteurs, le modèle *Big Five* [McCrae and John, 1992]. Ce modèle permet de comparer les traits de personnalité des individus en fonction de leur ouverture à l'expérience, leur caractère consciencieux, leur extraversion, leur agréabilité et leur névrosisme. Les systèmes de recommandation peuvent ainsi ensuite être adaptés pour fournir des recommandations plus ou moins diversifiées en fonction des traits de personnalité des utilisateurs.

Cependant, les différentes approches de diversification restent peu évaluées dans le contexte particulier de la recommandation de news. Le rôle démocratique de ces systèmes [Helberger, 2021], qui influencent la façon dont est mené le débat public, souligne pourtant l'importance d'un ap-

port contrôlé et personnalisé en diversité [Bernstein *et al.*, 2020]. Selon moi, **les systèmes de recommandation de news jouant un rôle dans le phénomène de polarisation, l'apport en diversité doit reposer sur une modélisation précise des besoins individuels en fonction des comportements de polarisation adoptés**, telle que celle proposée dans la Partie I de ce manuscrit.

4.3.2 Au-delà de la diversité

Dans une perspective de réduction du phénomène de polarisation, les concepteurs de systèmes de recommandation doivent également veiller à ne pas guider l'opinion des utilisateurs sous l'effet de biais algorithmiques. Au contraire, l'objectif est de faire prendre conscience aux utilisateurs de ce qui existe, sans les influencer. Les recommandations ne doivent donc pas être délibérément orientées en faveur de certaines opinions, au risque d'influencer les utilisateurs. Les recommandations ne doivent pas non plus sur-compenser les éléments qui sont mis de côté par les utilisateurs sous prétexte de diversification [Biega *et al.*, 2018]. En ce sens, le respect du concept de *calibration* [Steck, 2018], qui a attiré l'attention des chercheurs s'intéressant au principe d'équité dans les systèmes de recommandation [Wang *et al.*, 2023], est nécessaire. Pour garantir l'équité des recommandations, une liste de recommandations doit rester calibrée avec les différents intérêts d'un utilisateur. [da Silva *et al.*, 2021] ont mis en avant l'effet positif d'un calibrage personnalisé, tandis que [Abdollahpouri *et al.*, 2020] quantifient ce (mauvais) calibrage pour évaluer l'équité. Pour aller plus loin, [Kaya and Bridge, 2019a] proposent de comparer les recommandations calibrées et les recommandations diversifiées suivant une approche *intent-aware*. Ils concluent que les recommandations tenant compte de l'intention peuvent être calibrées et que les recommandations calibrées peuvent être diversifiées. La diversité et le calibrage sont donc étroitement liés.

Pour résumer, selon moi, la diversification des recommandations de news en vue de réduire le phénomène de polarisation ne peut être appliquée au détriment des dimensions éthiques. **L'apport en diversité doit reposer sur une approche personnalisée, respectueuse des intérêts de chaque utilisateur. Les dimensions d'exactitude, de diversité et d'équité sont donc toutes à prendre en compte, et aucune ne doit être artificiellement optimisée au détriment des autres.**

4.4 Synthèse et positionnement

Les systèmes de recommandation sont désormais omniprésents dans notre vie de tous les jours. Ils reposent sur une large variété d'approches, présentant toutes leurs avantages et leurs inconvénients. Leur développement a évolué parallèlement à leur utilisation croissante, passant d'une focalisation sur la précision à l'intégration de multiples critères tels que la diversité. L'éthique est également devenue une préoccupation majeure, assurant que les droits des utilisateurs soient respectés par ces systèmes.

En ce qui concerne spécifiquement la recommandation, au cœur de ce sujet de thèse, les approches de diversification apparaissent comme intuitives pour contrer, ou au moins limiter, leur impact sur le phénomène de polarisation [Stray, 2021]. Les stratégies de ré-ordonnancement *greedy* permettant de trouver un équilibre entre précision et diversité, bien qu'ayant prouvé leur efficacité, présentent néanmoins les limites suivantes :

— Elles ne permettent pas un apport personnalisé en diversité, pourtant primordial pour assurer un impact positif de l'apport en diversité sur les comportements de consommation

de news.

— Leur vision limitée ne permet pas de contrôler finement la nature de l'apport en diversité ou de l'équité.

— Elles ne sont pas adaptées à des situations impliquant l'optimisation de plus de deux critères.

Les approches personnalisées d'apport en diversité présentées précedemment répondent notamment aux deux premières limites. Elles restent cependant non étudiées dans le domaine spécifique de la recommandation de news, et ne permettent pas une optimisation multi-objectifs de plus de deux critères. Pourtant, en plus d'apporter de la diversité, ces approches visant à atténuer le phénomène de polarisation doivent également prendre en compte des considérations éthiques. Il est notamment crucial que la quête de diversification ne soit pas réalisée au détriment de l'utilisateur, en évitant toute forme d'influence et en respectant toujours ses besoins et préférences.

Pour résumer, **la littérature manque à ce jour de systèmes de recommandation de news capables de fournir des recommandations étant à la fois exactes, mais garantissant l'exposition à des points de vue divers, tout en étant équitables et respectant les préférences des utilisateurs** (Figure 4.4). Une telle approche serait d'un grand intérêt dans une perspective de réduction de la polarisation au travers d'une adaptation des recommandations. Cependant, très peu d'études proposent d'optimiser trois facteurs ou plus. Certaines approches peuvent pourtant être envisagées, comme le paradigme basé sur les contraintes [Felfernig *et al.*, 2015, Seymen *et al.*, 2021], approprié à la recommandation multi-objectifs [Jambor and Wang, 2010].

Les objectifs de travail permettant de répondre aux limites identifiées de la littérature, ainsi qu'à ma seconde question de recherche (*QR2*), sont donc les suivants :

— Proposer une approche de diversification personnalisée dans les recommandations de news

— Définir une approche de diversification garantissant à la fois exactitude, diversité et équité

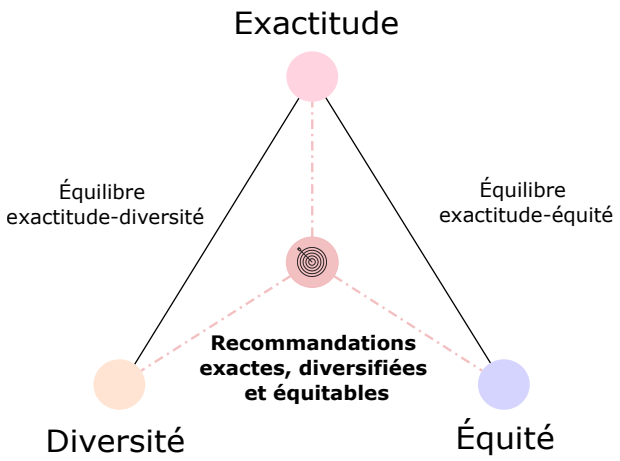


FIGURE 4.4 – Positionnement du travail - Recommandation

Chapitre 5

Diversification contrainte par l'équité : le framework ADF

Sommaire

5.1 Introduction

Pour répondre aux limites identifiées de la littérature relatives aux approches de diversification, à savoir leur manque de personnalisation et de contrôle de la diversité apportée ainsi que leur incapacité à optimiser plus de deux critères simultanément, je propose une nouvelle approche de diversification personnalisée. Cette approche est appliquée au domaine particulier de la recommandation de news. Elle a la particularité d'aller au-delà des traditionnelles approches bi-objectives existantes, et propose une optimisation multi-objectif de trois dimensions essentielles de la recommandation : (1) l'**exactitude**, assurant la satisfaction des utilisateurs et leur acceptabilité du système ; (2) la **diversité**, participant à la réduction du phénomène de bulles de filtre ;

(3) l'**équité**, permettant de s'abstraire de toute manipulation de l'opinion des utilisateurs. Ces trois dimensions sont communément optimisées dans la littérature, mais jamais simultanément. En lien avec ces objectifs, et pour compléter ma seconde question de recherche (*QR2*) portant sur la manière dont modifier les approches de recommandation de news de façon à apporter de la diversité tout en respectant les préférences des utilisateurs et les aspects éthiques, je pose deux sous-questions de recherche :

Sous-questions de recherche abordées dans ce chapitre :
Comment diversifier les recommandations tout en assurant leur équité ? (*QR2.1*)

Quel est l'impact de la diversification équitable sur l'exactitude des recommandations ? (*QR2.2*)

Dans la suite de ce chapitre, l'approche de diversification proposée, nommée ADF (Accuracy Diversity Fairness), est d'abord détaillée. Dans une seconde section, le protocole expérimental permettant d'évaluer cette approche est détaillé. Finalement, les résultats de cette validation expérimentale sont présentés.

5.2 ADF : un framework de recommandation multi-objectif

Les enjeux liés à la réduction du phénomène de polarisation au travers de l'adaptation des recommandations de news, au cœur de mon travail de thèse, résident dans la capacité à apporter de la diversité dans les recommandations, tout en assurant leur acceptabilité par les utilisateurs. Une stratégie permettant de répondre à cet enjeu, mise en avant dans les conclusions de la littérature et confirmée par la modélisation présentée dans la Partie I de ce manuscrit, consiste à apporter une diversité qui soit personnalisée, et adaptée au niveau de polarisation individuel de chacun des utilisateurs. Cependant, au-delà de cet apport en diversité, essentiel pour limiter la polarisation, une seconde dimension primordiale s'impose à la tâche de recommandation : l'équité. En effet, il est essentiel que la diversification des recommandations, bien qu'elle soit personnalisée, soit également contrôlée de façon à garantir que les avis ou opinions des utilisateurs ne soient pas artificiellement influencés ou modifiés sous prétexte de réduction de la polarisation.

Le framework de recommandation que je propose, appelé ADF pour *Accuracy-Diversity-Fairness*, est ainsi conçu pour promouvoir la sensibilisation des utilisateurs à l'information disponible, au travers de la diversité, tout en assurant de ne pas orienter leurs principaux intérêts ou opinions, au travers de l'équité. À ma connaissance, ADF est ainsi le premier framework qui cherche à optimiser la diversité et l'équité conjointement, tout en considérant l'exactitude des recommandations. Il consiste en une approche de ré-ordonnancement allant au-delà des compromis communément appliqués, tels qu'obtenus par l'approche *greedy*.

Comme détaillé dans l'état de l'art (Chapitre 4), le concept de diversité est complexe et les définitions sont multiples. Dans le domaine de la recommandation, et particulièrement de la recommandation de news, cette dimension de *diversité* est généralement évaluée sur l'ensemble des news N recommandées. Conformément à [Vargas, 2015], la *diversité* d'une liste de recommandations de news peut par exemple être évaluée sur un ensemble d'aspects A, caractérisant les news (genres, sujets, sources, *etc.*) [Steck, 2018, Kaya and Bridge, 2019a]. Ces aspects peuvent être définis à partir des méta-données disponibles, ou être définis selon un pré-traitement des

données exploitées. Selon [Vargas, 2015], la diversité permet ainsi d'*évaluer dans quelle mesure l'ensemble des aspects est diversement représenté dans une liste de news.*

De façon similaire, les définitions de l'équité dans la littérature sont multiples. Dans le travail présenté dans ce chapitre, et en accord avec les travaux de [Steck, 2018], l'*équité* est définie comme *la capacité à refléter les différents intérêts d'un utilisateur, en fonction de leurs proportions correspondantes.* De façon analogue à l'évaluation de la diversité, ces intérêts sont définis sur un ensemble d'aspects A caractérisant les news.

Ainsi, suivant ces définitions de diversité et d'équité, et étant donné un utilisateur u, le framework ADF s'appuie sur quatre notions clés :

1. La distribution des news sélectionnées par u sur les différents aspects (*c.-à-d.* les intérêts de u), appelée la ***distribution de sélection***.
2. La diversité associée à la ***distribution de sélection*** de u, appelée la ***diversité de sélection***.
3. La diversité attendue dans les recommandations faites à u suivant sa *diversité de sélection*, appelée ***diversité cible***.
4. La distribution associée à cette *diversité cible*, correspondant à la distribution des news recommandées sur les différents aspects, appelée ***distribution cible***.

Pour évaluer ces distributions et diversités puis calculer les recommandations leur correspondant, trois types d'informations sont fournies en entrée du framework ADF : (1) l'ensemble des aspects A permettant de caractériser les news et sur lesquels reposent les définitions de diversité et d'équité ; (2) les données d'interaction des utilisateurs, correspondant aux news sélectionnées par chaque utilisateur $u \in U$; et (3) l'ensemble des news non lues pour chaque utilisateur $u \in U$, associées à des scores de pertinence personnalisés fournis par un système de recommandation $R(u) = \{n, s(u,n)\}$, avec $n \in N$. Le score de pertinence $s(u,n)$ représente la mesure dans laquelle une news n correspond aux intérêts d'un utilisateur u. Ces scores sont exploités pour assurer l'exactitude des recommandations finales puisqu'ils permettent d'identifier les news susceptibles de correspondre aux attentes de u.

Je précise ici qu'**ADF est agnostique à l'algorithme de recommandation utilisé**, de sorte que n'importe quel algorithme de la littérature peut être appliqué, qu'il soit basé sur le contenu ou collaboratif.

À partir de ces données fournies en entrée, le framework de recommandation proposé ADF est décomposé en 5 étapes (Figure 5.1). Chacune de ces étapes est détaillée dans les sous-sections suivantes. Pour guider la lecture et illustrer le fonctionnement du framework, sa description est ponctuée d'un exemple fil rouge, indépendant de celui présenté dans le Chapitre 2. Le Tableau 5.1 résume les différentes notations utilisées dans l'ensemble de ce chapitre.

TABLE 5.1 – Notations utilisées dans la suite du manuscrit

u	un utilisateur	U	ensemble d'utilisateurs	n	une news
N	ensemble de news	$\|N\|$	nombre de news	a	un aspect
A	ensemble d'aspects	$\|A\|$	nombre d'aspects	$P(u)$	distribution de sélection de u
$p(a\|u)$	probabilité de sélection de u pour a	$P^*(u)$	distribution cible de u	$p^*(a\|u)$	probabilité cible de u pour a
s(u,n)	score de pertinence de u pour n	$R(u)$	liste de news de u & scores de pertinence	$R^*_k(u)$	liste de recommandations de u (top-k news & scores de pertinence)

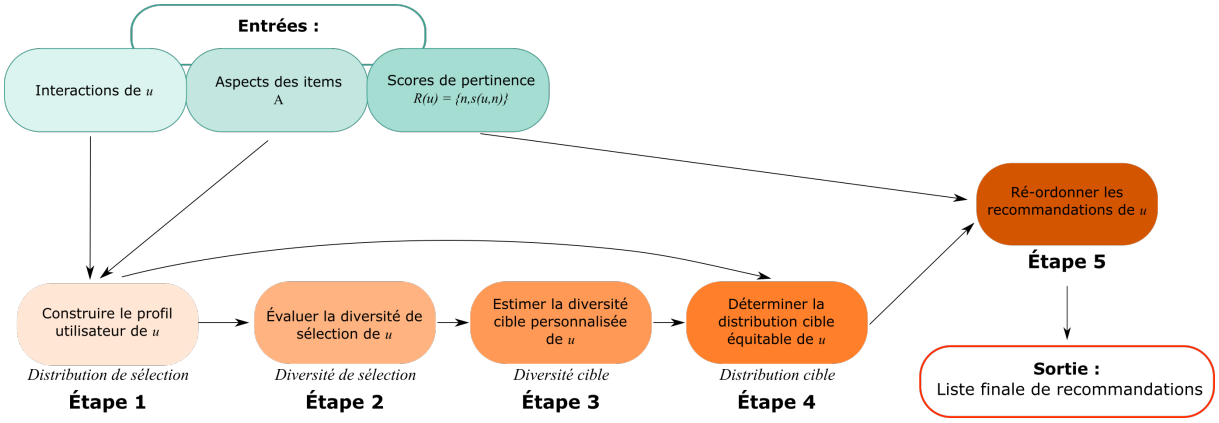


FIGURE 5.1 – Étapes du framework ADF

5.2.1 Étape 1 : construire le profil utilisateur de u

La première étape du framework ADF est consacrée à la **construction du profil utilisateur**, correspondant à la représentation des intérêts d'un utilisateur u sur l'ensemble des aspects A, à partir de ses interactions sur les news. Concrètement, l'intérêt de u pour un aspect donné $a \in A$ est formulé comme une probabilité $p(a|u)$. Cette probabilité $p(a|u)$ correspond à la probabilité de l'aspect a dans les interactions de u [Eskandanian *et al.*, 2017]. Un profil utilisateur final est ainsi une distribution de probabilité $P(u)$ sur l'ensemble des aspects, avec $\sum_{a \in A} p(a|u) = 1$. Cette distribution $P(u)$ correspond à l'une des quatre notions clés sur lesquelles repose ADF : la *distribution de sélection*.

Cette *distribution de sélection* est exploitée pour optimiser deux dimensions au cœur du framework ADF. D'une part, en représentant les intérêts d'un utilisateur u, elle participe à assurer l'**équité** en suivant l'approche proposée par [Steck, 2018]. D'autre part, elle peut être exploitée de façon à adapter l'apport en **diversité**. Cette double utilité de la *distribution de sélection* contribue à l'originalité du framework proposé. Par ailleurs, la *distribution de sélection* est propre à chaque utilisateur, ce qui contribue également à personnaliser l'approche de diversification des recommandations.

> J'introduis ici l'exemple fil rouge, qui sera repris et complété dans la suite de ce chapitre. Soit u_1 un utilisateur qui interagit avec un ensemble N de news représentées dans l'espace d'aspects A correspondant aux sujets des news, avec $|A| = 6$ aspects. Le profil de u_1 peut alors être construit en fonction de la proportion de ses interactions avec chaque aspect $a \in A$. La distribution résultante $P(u_1)$ correspond à la *distribution de sélection* de u_1, dont un exemple est présenté dans la Figure 5.2. Elle permet d'identifier que l'utilisateur u_1 a un grand intérêt pour l'aspect a_2, un intérêt modéré pour l'aspect a_3, le même intérêt pour les aspects a_4 et a_6, aucun intérêt pour l'aspect a_5, *etc.*

5.2.2 Étape 2 : évaluer la diversité de sélection de u

Étant donné la *distribution de sélection* $P(u)$, établie lors de l'étape 1, **l'objectif de la deuxième étape est d'évaluer la diversité des intérêts de** u, notée $diversity(P(u))$, c'est-à-dire la diversité associée à cette distribution. La diversité étant un concept très large et associé à

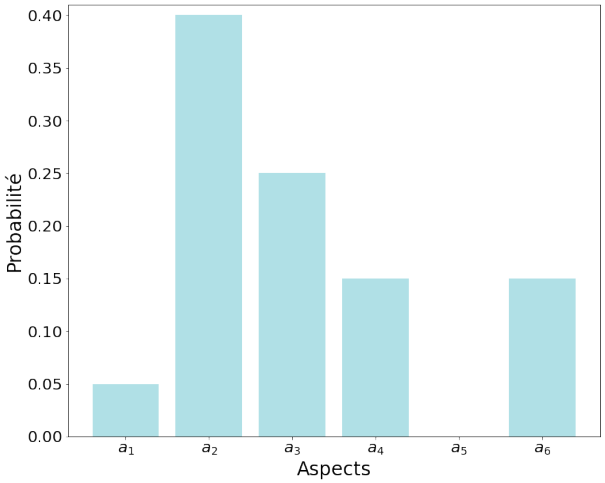


FIGURE 5.2 – *Distribution de sélection* de u_1

de multiples définitions, plusieurs mesures issues de la littérature peuvent être utilisées pour l'instancier. En particulier, dans mon cadre de recherche sur la polarisation, il est possible d'appliquer des métriques telles que le score de polarisation [Becatti *et al.*, 2019], l'entropie [Shannon, 1948], *etc.* Néanmoins, pour l'application du framework ADF, comme pour toute mesure de diversité et tel que proposé dans la Chapitre 2, il est préférable que $\forall u \in U,\ diversity(P(u)) \in [0;1]$ afin d'assurer la comparabilité entre les utilisateurs et permettre une meilleure interprétation.

Pour revenir à l'exemple fil rouge, la diversité de sélection de u_1 peut être évaluée par l'entropie normalisée (Équation 2.1), avec $diversity(P(u_1)) = 0,80$.

5.2.3 Étape 3 : estimer la diversité cible personnalisée de u

Avant de détailler cette troisième étape du framework ADF, je rappelle qu'un fort apport en diversité dans les recommandations ne contribue pas systématiquement à une consommation plus diverse de l'information de la part des utilisateurs. Au contraire, certains utilisateurs peuvent même recevoir négativement cette diversification qui peut être qualifiée d'abusive, en particulier lorsque la liste de recommandations entre en conflit avec leurs préférences et opinions. Dans ce cas, la diversité peut renforcer la polarisation politique de l'utilisateur [Bail *et al.*, 2018]. Ces résultats confirment ainsi la nécessité d'un niveau personnalisé de diversité des recommandations [Treuillier *et al.*, 2022]. En complément de ces conclusions de l'état de l'art, les travaux de modélisation détaillés dans la première partie de ce manuscrit ont permis de discriminer des classes de comportements de polarisation. Ces dernières se distinguent notamment par leur niveau de polarisation, auquel il semble alors primordial d'adapter les stratégies de diversification dans les recommandations. **L'objectif de cette troisième étape est donc d'estimer le niveau de diversité personnalisé attendu dans les recommandations d'un utilisateur u.** Cette diversité est appelée *diversité cible*, et est notée $diversity(P^*(u))$, où $P^*(u)$ est la *distribution cible* attendue dans les recommandations fournies à un utilisateur u.

Pour définir cette *diversité cible* personnalisée, l'approche que je propose consiste à exploiter la *diversité de sélection* d'un utilisateur u, évaluée lors de l'étape 2 (Section 5.2.2). Ainsi, la *diversité cible* est définie comme une fonction f de la *diversité de sélection*, $f(diversity(P(u)))$. Fondamentalement, la *diversité cible* peut être égale à la *diversité de sélection*, avec $f(x) = x$ où

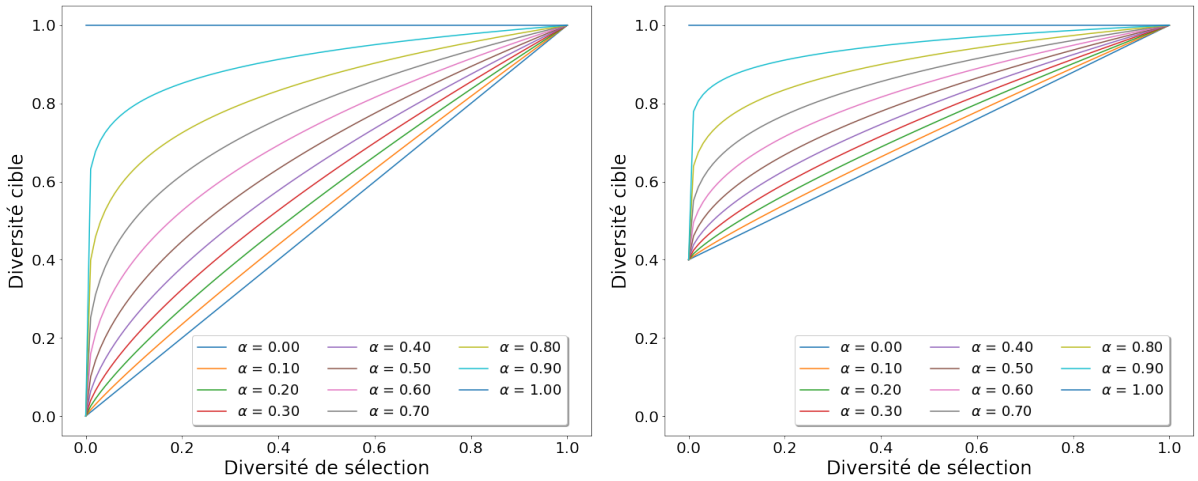
x est une valeur de diversité. L'objectif étant d'apporter de la diversité dans les recommandations, et de promouvoir la sensibilisation de u à l'ensemble des aspects A, deux caractéristiques principales sont attendues pour la fonction $f()$:

1. La *diversité cible* d'un utilisateur ne peut être inférieure à sa *diversité de sélection*. Concrètement, cela signifie que les utilisateurs ne recevront pas de recommandations dont la diversité est inférieure à leur propre diversité de sélection. Ainsi, $\forall x, x \in [0;1], f(x) \geq x$.
2. Si un utilisateur u_1 a une *diversité de sélection* supérieure à celle d'un utilisateur u_2, la *diversité cible* de u_1 sera supérieure ou égale à la *diversité cible* de u_2. Concrètement, la fonction f doit être croissante. Ainsi,
   $diversity(P(u_1)) > diversity(P(u_2)) \Rightarrow f(diversity(P(u_1))) \geq f(diversity(P(u_2)))$.

La formulation de la fonction f proposée et répondant à ces deux caractéristiques est donnée dans l'Équation 5.1.

$$f(x) = \beta + (1 - \beta)x^{1-\alpha} \tag{5.1}$$

avec $0 \leq \alpha, \beta \leq 1$. Le paramètre β permet de définir la diversité minimale des diversité cibles, tandis que le paramètre α permet de faire varier l'inclinaison et la pente de la courbe. Lorsque $\alpha = 0$, la fonction $f()$ est linéaire et l'augmentation de la diversité est constante. Dans ce cas, si $\beta = 0$, cela revient à $f(x) = x$. Cependant, définir une *diversité cible* égale à la *diversité de sélection* ne contribue pas à construire une liste de recommandations plus diversifiée, ce qui est limitant pour les utilisateurs ayant une faible *diversité de sélection*. A l'inverse, lorsque $\alpha = 1$, la *diversité cible* est maximale (égale à 1), et est indépendante de la *diversité de sélection*. Dans ce cas, la valeur de β n'a aucun impact sur la *diversité cible*. Ainsi, plus la valeur de α est élevée, plus la *diversité cible* est élevée. La Figure 5.3 représente l'évolution de la fonction f avec plusieurs valeurs de α, lorsque $\beta = 0$ (Figure 5.3a) et lorsque $\beta = 0,4$ (Figure 5.3b).



(a) α variant dans $[0;1]$, $\beta = 0$.

(b) α variant dans $[0;1]$, $\beta = 0,4$.

FIGURE 5.3 – Fonction de diversification $f()$ avec différentes valeurs des paramètres α et β.

Si je reprends l'exemple fil rouge, l'application de la fonction f avec $\alpha = 0,4$, et $\beta = 0$, alors $f(diversity(P(u_1))) = 0,87$. La *diversité cible* pour l'utilisateur u_1 est donc de $0,87$.

5.2.4 Étape 4 : déterminer la distribution cible équitable de u

Étant donnée la *diversité cible* de u estimée à l'étape 3, **l'objectif de cette quatrième étape est de déterminer la *distribution cible* dont la diversité est égale à cette *diversité cible*.** Il s'agit donc de déterminer $P^*(u) = \{p^*(a|u)\}, a \in A$, avec $diversity(P^*(u)) = f(diversity(P(u))$. Cette distribution cible est celle qui sera fournie dans les recommandations.

Cependant, je rappelle qu'une caractéristique essentielle du framework ADF est d'assurer l'équité des recommandations. La définition de l'équité sur laquelle s'appuie le framework, énoncée dans l'introduction de chapitre, repose sur les intérêts d'un utilisateur et leurs proportions. En d'autres termes, **le framework ADF considère qu'une liste de recommandations est équitable pour un utilisateur u, si la *distribution cible* sur l'ensemble des aspects A est compatible avec la *distribution de sélection* de u sur A.** Pour assurer cette compatibilité entre la distribution de sélection et la distribution cible, **une contrainte est imposée par le framework ADF**.

Compatibilité des distributions : notion de contrainte

La distribution dont la diversité égale une *diversité cible* n'est pas unique. En effet, des distributions de probabilité différentes peuvent avoir la même diversité, quelle que soit la manière dont la diversité est évaluée. Ainsi, l'objectif de cette étape n'est pas uniquement d'identifier une *distribution cible* $P^*(u)$ dont la diversité est égale à la *diversité cible*, mais surtout d'identifier cette *distribution cible* qui soit compatible avec la *distribution de sélection* $P(u)$ pour garantir l'équité des recommandations. Ceci contraint la *distribution cible* et la rend unique.

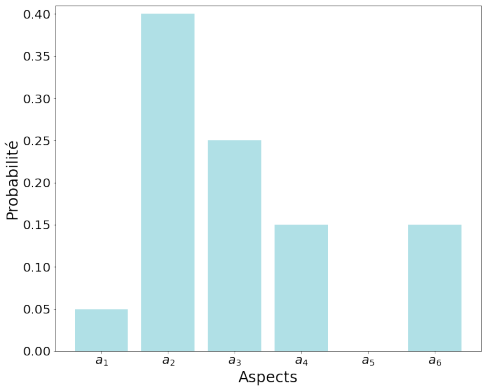
L'originalité de cette approche, consistant à définir la distribution des recommandations sur différents aspects de façon à apporter de la diversité, est double. Elle réside premièrement dans l'optimisation conjointe de la diversité et de l'équité, qui n'a, à ma connaissance, jamais été abordée dans la littérature. Deuxièmement, elle réside dans la manière dont l'optimisation des deux dimensions est définie. Dans la littérature, l'optimisation précision-équité ou précision-diversité est généralement définie par une fonction permettant de faire varier le poids de chacune des dimensions. Il en résulte un compromis entre les deux dimensions, qui ne permet pas de garantir l'optimisation des deux dimensions. Cependant, dans le framework ADF, je propose d'imposer l'optimisation de l'équité, qui prime sur l'exactitude des recommandations. Ainsi, la *distribution de sélection* agit comme une contrainte sur la *distribution cible*, permettant d'optimiser la diversité et l'équité de façon simultanée. Concrètement, il s'agit d'une contrainte de compatibilité, utilisée pour garantir l'équité de la *distribution cible*, sur laquelle repose l'apport en diversité des recommandations.

Cette contrainte de compatibilité ne peut être simplement définie comme une équivalence entre les deux distributions, car elle ne permettrait pas d'estimer une distribution avec une *diversité cible* plus élevée. La compatibilité doit permettre des différences entre les deux distributions. Pour répondre à cela, je propose d'instancier la compatibilité en exploitant les relations d'ordre des distributions. Concrètement, deux distributions de probabilités sont compatibles si leurs relations d'ordre sont égales, *c.-à-d.* que l'ordre entre les éléments est le même dans les deux distributions. En d'autres termes,

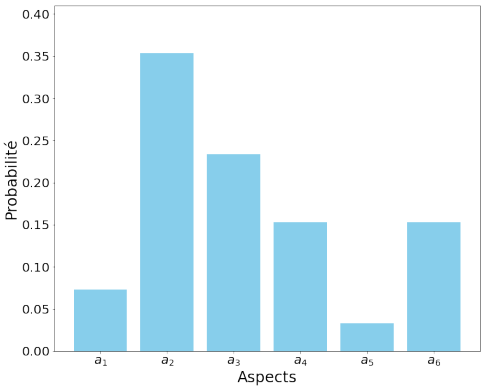
$$\forall a_i, a_j \in A \begin{cases} \texttt{si} \quad p(a_i|u) > p(a_j|u) \quad \texttt{alors} \quad p^*(a_i|u) \geq p^*(a_j|u) \\ \texttt{si} \quad p(a_i|u) < p(a_j|u) \quad \texttt{alors} \quad p^*(a_i|u) \leq p^*(a_j|u) \\ \texttt{si} \quad p(a_i|u) = p(a_j|u) \quad \texttt{alors} \quad p^*(a_i|u) = p^*(a_j|u) \end{cases} \tag{5.2}$$

avec $p(a_i|u)$ la *probabilité de sélection* de l'utilisateur u pour l'aspect a_i, et $p^*(a_i|u)$ la *probabilité cible* de l'utilisateur u pour l'aspect a_i.

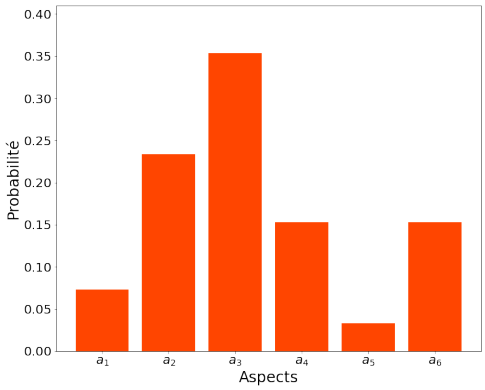
La figure 5.4a illustre cette notion de compatibilité à partir de $P(u_1)$, la *distribution de sélection* de l'utilisateur exemple u_1. La distribution $P'(u_1)$ de la Figure 5.4b est compatible avec $P(u_1)$. Cependant, la distribution $P''(u_1)$ de la Figure 5.4c a la même valeur de diversité (entropie normalisée) que $P'(u_1)$, mais est incompatible avec $P(u_1)$. En effet, $p(a_2|u_1) > p(a_3|u_1)$ mais $p''(a_2|u_1) < p''(a_3|u_1)$. Suivant la définition de l'équité sur laquelle repose ADF, cette distribution n'est donc pas incompatible avec la distribution de sélection de u_1, et les recommandations résultantes seraient ainsi non équitables pour u_1.



(a) *Distribution de sélection* $P(u_1)$



(b) Distribution compatible $P'(u_1)$



(c) Distribution incompatible $P''(u_1)$

FIGURE 5.4 – Distributions compatible et incompatible avec la *distribution de sélection* de u_1

Maintenant que la notion de contrainte est introduite, il s'agit de définir une fonction de transformation équitable d'une distribution, *c.-à-d.* qui permet de former une distribution compatible afin de garantir l'équité des recommandations.

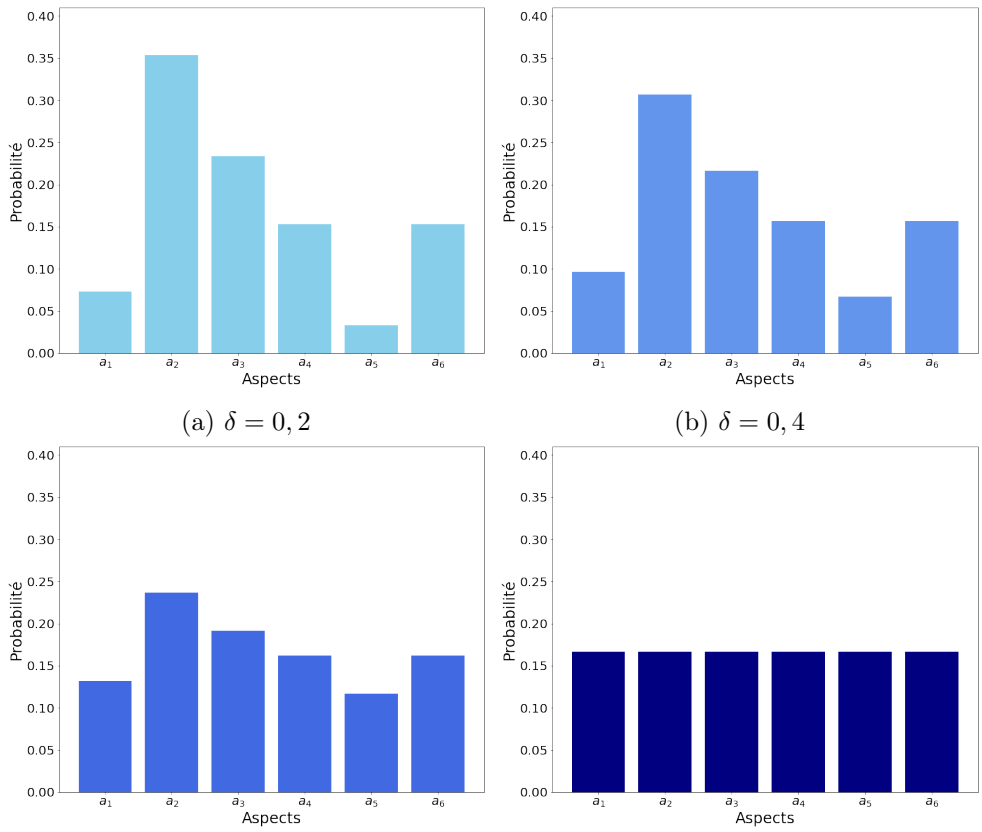
Transformation équitable d'une distribution de probabilité

Je laisse ici temporairement de côté l'identification de la *diversité cible*, et me concentre sur la façon de transformer une distribution $P(u)$ en une distribution compatible $P^*(u)$, de façon à ce que $P^*(u)$ respecte les caractéristiques formulées dans l'Équation (5.2). Soit *smooth*() cette fonction de transformation, présentée dans l'Équation (5.3).

$$smooth(u, \delta) = (1 - \delta)P(u) + \delta\frac{1}{|A|} \tag{5.3}$$

où $\frac{1}{|A|}$ est la distribution de probabilité uniforme et le paramètre δ, $0 \leq \delta \leq 1$, représente son poids. Cette fonction de transformation s'applique à chaque aspect $a \in A$, où $P(u)$ est instancié par $p(a|u)$. Concrètement, *smooth*() est une combinaison linéaire entre la *distribution de sélection* ($P(u)$) et la distribution de probabilité uniforme ($\frac{1}{|A|}$). Comme la transformation est appliquée à chaque élément de la distribution de probabilité, les contraintes de l'Équation (5.2) sont remplies, la distribution est lissée et la distribution résultante est compatible, et donc équitable. L'intensité de la transformation dépend du paramètre δ. Si $\delta = 0$, aucun lissage n'est effectué et la distribution est inchangée. Au contraire, si $\delta = 1$, la distribution résultante est totalement uniformisée. Ainsi, plus la valeur de δ est élevée, plus la fonction uniforme a de poids et donc plus la distribution est lissée.

La figure 5.5 présente la *distribution de sélection* transformée de l'utilisateur exemple u_1, avec différentes valeurs de δ.



(a) $\delta = 0,2$ (b) $\delta = 0,4$

(c) $\delta = 0,7$ (d) $\delta = 1$

FIGURE 5.5 – Application de la fonction *smooth*() sur la *distribution de sélection* de u_1 avec différentes valeurs de δ.

Il est important de noter que les aspects ayant une probabilité nulle sur la *distribution de sélection* ont une probabilité non nulle dans la *distribution cible*, ce qui contribue à apporter de

la diversité. Néanmoins, ces aspects restent ceux dont la valeur de probabilité est la plus faible, ce qui est conforme à la notion de compatibilité introduite, et permet donc d'assurer l'équité.

À partir de la fonction *smooth*() permettant de transformer une distribution en une distribution compatible, il s'agit finalement d'identifier le paramètre optimal δ permettant d'obtenir une distribution transformée dont la diversité associée est au plus proche de la diversité cible définie dans l'Étape 3 (Section 5.2.3).

Détermination du paramètre de transformation δ

Pour finaliser cette quatrième étape, l'objectif est donc de déterminer, pour un utilisateur donné u, la valeur δ qui transforme la *distribution de sélection* $P(u)$ de sorte que la diversité associée à la distribution transformée se rapproche de la *diversité cible*, $f(diversity(P(u)))$. Ce problème correspond à une optimisation mono-objectif, présentée dans l'Équation (5.4).

$$\delta^* = \underset{\delta}{\operatorname{argmin}} \ \left| f(diversity(P(u))) - diversity(smooth(P(u), \delta)) \right| \tag{5.4}$$

Par conséquent, avec δ^* la valeur optimale de δ, $smooth(P(u), \delta^*)$ correspond à la distribution dont la diversité associée est la plus proche de la *diversité cible*. Cette distribution correspond ainsi à la *distribution cible* de l'utilisateur u, notée $P^*(u)$.

Pour revenir à l'exemple fil rouge, sur la base de la *distribution de sélection* de u_1, la *diversité de sélection* et la *diversité cible* associée sont calculées. La valeur optimale de δ est $\delta^* = 0,2$, ce qui permet d'obtenir la *distribution cible*. Les quatre premières étapes du framework ADF, depuis la détermination de la *distribution de sélection* de u_1 jusqu'à la détermination de sa *distribution cible*, sont résumées dans la Figure 5.6.

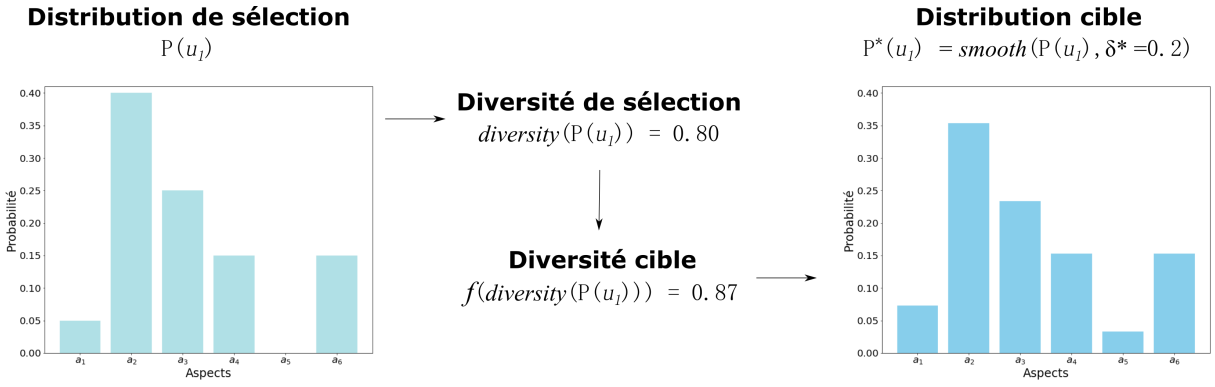


FIGURE 5.6 – Détermination de la *distribution cible*

5.2.5 Étape 5 : ré-ordonner les recommandations de u

L'objectif de cette cinquième et dernière étape est de ré-ordonner les recommandations suivant la distribution cible, permettant d'apporter le niveau de diversité cible. Concrètement, cette dernière étape consiste à ré-ordonner $R(u)$, la liste de scores de pertinence fournie en entrée d'ADF, en une liste de recommandations finale $R_k^*(u)$ de longueur k.

Les stratégies de ré-ordonnancement de la littérature sont principalement basées sur l'approche (*greedy*) [Vargas and Castells, 2013, Ziegler *et al.*, 2005]. Cependant, de part son fonctionnement (détaillé dans l'état de l'art, Chapitre 4), une telle approche ne peut permettre de garantir le respect de l'équité dans les recommandations, ce qui va à l'encontre de l'objectif annoncé du framework ADF. Le ré-ordonnancement des recommandations est donc effectué de manière à ce que la liste de recommandations résultante $R_k^*(u)$ respecte la contrainte d'équité, tout en maximisant la précision. Ici encore, l'équité agit comme une contrainte et la distribution des aspects de $R_k^*(u)$ doit être au plus proche de la *distribution cible* $P^*(u)$.

La stratégie de ré-ordonnancement proposée est appliquée à partir de trois entrées :

1. $R(u)$ la liste des news avec les scores de pertinence,
2. $P^*(u)$ la *distribution cible* personnalisée,
3. k la taille de la liste de recommandations attendue.

A partir de ces éléments, la stratégie de ré-ordonnancement fonctionne en deux étapes. Premièrement, le nombre attendu de news par aspect $a \in A$ que la liste finale de recommandation doit contenir pour que la *distribution cible* ($P^*(u)$) soit respectée est calculé. Étant donné un aspect a, ce nombre, noté à l_a, est calculé ainsi : $l_a = round(p^*(a|u) \cdot k)$. Deuxièmement, les news qui sont étiquetées avec l'aspect a et dont les scores de pertinence les plus élevés dans la liste initiale $R(u)$, notées top-l_a, sont sélectionnées. La sélection des news avec les scores de pertinence les plus élevés permet de maximiser l'exactitude, une dimension clé de la tâche de recommandation.

Le résultat de cette stratégie de reclassement est donc la liste de recommandations finale $R_k^*(u)$, composée des top-l_a news pour chaque aspect a, et dont la distribution est aussi proche que possible de la *distribution cible* personnalisée $P^*(u)$.

> Pour finaliser l'exemple fil rouge, la liste de recommandations $R_{20}^*(u_1)$ contient 20 recommandations dont la distribution sur les aspects A est la plus proche de la *distribution cible* $P^*(u_1)$. L'aspect le plus représenté dans cette liste finale de recommandations est l'aspect a_2, qui est celui sur lequel u_1 porte le plus grand intérêt selon sa distribution de sélection $P(u_1)$. Les autres aspects sont néanmoins tous présents dans la liste, et notamment l'aspect a_5, pour lequel l'utilisateur u_1 n'avait initialement aucun intérêt. Cette liste finale de recommandations $R_{20}^*(u_1)$ est ainsi diversifiée et équitable car la distribution des aspects est compatible avec celles des intérêts de u_1.

5.3 Protocole d'évaluation

5.3.1 Objectif

L'évaluation du framework de recommandation ADF présentée dans ce chapitre a pour objectif d'évaluer dans quelle mesure il est possible d'optimiser simultanément les dimensions d'exactitude, de diversité et d'équité dans le cadre d'une tâche de recommandation. La sous-question à laquelle je cherche à répondre ici est donc (*QR2.2*) : Quel est l'impact de la diversification équitable sur l'exactitude des recommandations ?

5.3.2 Description du protocole

Pour répondre à cet objectif, je propose un protocole d'évaluation permettant premièrement d'évaluer les performances de recommandation lorsque ces dernières sont ré-ordonnées avec le

framework ADF ou avec des approches baseline issues de la littérature. Cette première phase du protocole permet d'évaluer dans quelle mesure l'optimisation multi-objectif appliquée avec ADF impacte les performances. Deuxièmement, l'apport en diversité permis par le framework ADF étant personnalisé, la seconde phase du protocole consiste à évaluer dans quelle mesure cette personnalisation impacte les performances par rapport à un apport en diversité global. Un schéma récapitulatif du protocole expérimental est présenté dans la Figure 5.7.

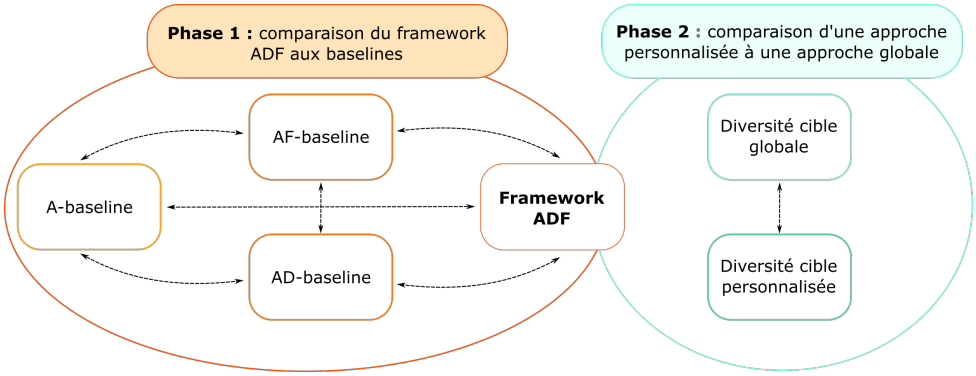


FIGURE 5.7 – Protocole expérimental pour l'évaluation du framework ADF

Phase 1 : comparaison du framework ADF aux baselines

Pour cette première phase du protocole, les performances du framework ADF sont comparées à trois modèles de référence, appelés *baselines* :

1. La **A-baseline** (**A**ccuracy), qui optimise uniquement la dimension d'exactitude. La liste de recommandations associée est $R_k(u)$, *c.-à-d.* la liste de k news avec le top-k des news ayant les scores de pertinence $s(u,n)$ les plus élevés dans $R(u)$.
2. La **AF-baseline** (**A**ccuracy-**F**airness), qui optimise à la fois la dimension d'exactitude et la dimension d'équité, et fournit une liste de recommandations parfaitement équilibrée. Cette baseline correspond à l'application du framework ADF avec le paramètre α fixé à 0, donc la *diversité cible* est égale à la *diversité de sélection*.
3. La **AD-baseline** (**A**ccuracy-**D**iversity), qui optimise à la fois la dimension d'exactitude et la dimension de diversité, en appliquant un ré-ordonnancement *greedy*, tel que communément appliqué dans la littérature, suivant l'Équation (4.14) et l'Algorithme (1). Lors de l'application de cette approche *greedy*, les valeurs du paramètre $\lambda \in [0;1]$ avec $\Delta\lambda = 0,1$ sont testées. Pour rappel, ce paramètre permet d'ajuster l'équilibre entre l'exactitude et la diversité.

Ainsi, pour cette validation expérimentale, l'A-baseline est dans un premier temps comparée à l'AF-baseline afin d'étudier l'impact de la dimension d'équité. Dans un second temps, la contribution de la dimension de diversité est étudiée au travers de la comparaison entre la AD-baseline et la A-baseline. Dans un troisième temps, l'apport de ADF, qui optimise l'ensemble des trois dimensions est évalué, en comparant les résultats à chacune des trois baselines. Pour cette évaluation ADF est appliqué avec les valeurs de $\alpha \in [0;1], \Delta\alpha = 0,1$. Pour rappel, plus α est proche de 1, plus la diversité cible est élevée. Par ailleurs, la *diversité de sélection* moyenne dans le jeu de données étudié étant relativement élevée $(0,653)$, et aucun utilisateur n'ayant une diversité nulle $(min(diversit(P(u)) = 0,11)$, le paramètre β est fixé à 0.

Phase 2 : comparaison d'une approche personnalisée à une approche globale

Cette seconde phase du protocole propose d'évaluer dans quelle mesure un apport en diversité personnalisé, que je défends dans cette thèse, permet de contrôler l'apport en diversité dans les recommandations, et de maintenir leur exactitude. En d'autres termes, je propose de mesurer l'apport d'une diversification personnalisée telle que proposée par le framework ADF avec l'estimation d'une diversité cible propre à chaque utilisateur (Étape 3, Section 5.2.3). Cette diversification personnalisée est ainsi comparée à une diversification globale, correspondant à l'estimation d'une diversité cible commune pour l'ensemble des utilisateurs. Pour cette évaluation, lors de l'Étape 3 du framework ADF (Section 5.2.3), une valeur de diversité cible (*c.-à-d.* l'entropie) comprise entre $[0,6; 1]$, avec $\Delta = 0,1$, est fixée pour tous les utilisateurs. Les étapes 4 et 5 de ADF sont ensuite appliquées sans modification particulière (Sections 5.2.4 et 5.2.5 respectivement). Les performances de recommandation obtenues lorsque la diversité est personnalisée sont ensuite comparées à celles obtenues lorsque la diversité est globale.

5.4 Validation expérimentale du framework ADF

Cette section débute par une présentation du contexte sur lequel a été évalué le framework ADF. La configuration expérimentale est ensuite présentée. L'ensemble des résultats issus de la validation expérimentale du framework ADF est finalement présenté.

5.4.1 Contexte applicatif : l'agrégateur Microsoft News

Comme détaillé dans l'état de l'art (Chapitre 4), les jeux de données de référence pour la recommandation de news (Tableau 4.2) présentent des spécificités sur le type de données disponibles, la langue de rédaction des news, la période couverte, *etc.* Parmi ces jeux de données, celui qui me semblait le plus adapté pour évaluer le framework ADF est le jeu de données MIND (MIcrosoft News Dataset) [Wu *et al.*, 2020b]. En effet, celui-ci est particulièrement intéressant car il couvre une période de collecte des données étendue (6 semaines) par rapport aux autres jeux de données. De plus, il contient des news rédigées en anglais pour lesquelles les modèles de traitement automatique des langues existants sont plus adaptés par rapport à des contenus rédigés dans d'autres langues (norvégien pour le jeu de données Adressa, allemand pour le jeu de données Plista par exemple). Enfin, ce jeu de données est le plus largement utilisé dans le domaine de la recommandation de news [Vrijenhoek, 2023].

Le jeu de données MIND contient les interactions d'1 million d'utilisateurs sur les news publiées sur le site Microsoft News [18] entre le 12 octobre et le 22 novembre 2019. Ces utilisateurs ont été échantillonnés aléatoirement et sont anonymisés. Les données d'interaction sont enregistrées sous la forme de *logs d'impression*, contenant chacun un id utilisateur, l'historique de consommation de l'utilisateur au moment du *log*, la liste des news affichées sur l'écran de l'utilisateur, et les news avec lesquelles ce dernier a interagi. Au total, ce sont 15 millions d'impressions, à propos de 160 000 news qui ont été collectées. A propos de chaque news, le jeu de données fournit les informations suivantes : catégorie, sous-catégorie, titre, résumé et URL.

Le jeu de données MIND est divisé en trois sous-ensembles : l'ensemble d'apprentissage (*train set*), l'ensemble de test (*test set*), et l'ensemble de validation (*validation set*). Ainsi, les interactions produites durant les quatre premières semaines de collecte (entre le 12 octobre et le 8 novembre 2019) ont permis de définir l'historique de chaque utilisateur, tandis que les données des $5^{ème}$ et $6^{ème}$ semaines (entre le 9 novembre et le 22 novembre 2019) ont permis de construire

18. https://news.microsoft.com/source/

ces ensembles d'apprentissage, de test et de validation. Une version réduite du jeu de données est disponible (MIND-*small*), et correspond à une sélection aléatoire de 50 000 utilisateurs dans le jeu de données initial (MIND-*large*), et ne couvrant que les cinq premières semaines de collecte.

Une expérimentation préliminaire visant à évaluer la diversité globale des recommandations dans le jeu de données MIND, menée en amont de ce travail sur l'apport en diversité contraint par l'équité, est présentée dans l'Annexe B.

5.4.2 Éléments de configuration du framework ADF

Cette sous-section détaille les différents choix de configuration qui ont été appliqués pour l'évaluation du framework ADF sur le jeu de données MIND.

Sélection des données

Le framework ADF est évalué sur la version large du jeu de données MIND-*large*. Parmi l'ensemble des interactions disponibles, seules les interactions portant sur les news étiquetées de la catégorie "*news*" sont étudiées. En effet, les news associées à cette catégorie dans le jeu de données MIND sont celles se rapprochant le plus de news traitant de sujets politiques, et sont les plus susceptibles d'impacter le phénomène de polarisation politique, qui est au cœur de ma thèse.

Pour la tâche de recommandation en elle-même, seuls les utilisateurs ayant produit entre 20 et 200 interactions sur la période de collecte des données ont été sélectionnés. La borne inférieure permet d'assurer une quantité de données suffisante pour fournir des recommandations de qualité, tandis que la borne supérieure permet de filtrer les utilisateurs adoptant un comportement extrême, non représentatif de la population.

Enfin, afin de valider la pertinence du framework proposé et de fournir une preuve de concept, un sous-ensemble de 10 000 utilisateurs a été sélectionné. Ce choix est notamment motivé par le besoin d'optimiser les performances computationnelles, en permettant de réduire les coûts par rapport à l'application du modèle sur un ensemble de données plus volumineux, tout en conservant une population suffisamment large pour obtenir des résultats pertinents. **Le framework ADF a donc été évalué sur un ensemble de 10 000 utilisateurs, ayant interagi avec 18 186 news distinctes**.

Représentation numérique du contenu des news

Le corps du texte, qui est notamment nécessaire à l'application d'une approche de recommandation basée contenu, largement appliquée dans le domaine de la recommandation de news, n'est pas fourni dans le jeu de données MIND. Cependant, en exploitant l'URL fourni pour chacune des news, l'intégralité du contenu de chaque news peut être extrait automatiquement à l'aide d'un programme de *scraping*.

Une fois collecté, le contenu des news est représenté numériquement en appliquant l'approche LDA (Latent Dirichlet Allocation) [Blei *et al.*, 2003]. Cette méthode issue du domaine du traitement automatique des langues permet de représenter des contenus textuels sous forme de vecteurs particuliers, appelés *embeddings*. Appliqué à l'ensemble des news, cette méthode permet donc d'identifier les thèmes sous-jacents, puis de représenter chaque document comme une distribution sur ces thèmes. Les *embeddings* résultants comportent ainsi autant de dimensions que de thèmes, et correspondent à la distribution de probabilité sur chaque thème pour un document. La méthodologie complète d'application de LDA au jeu de données MIND est présentée dans l'Annexe C.

Algorithme de recommandation

Comme détaillé lors de la présentation d'ADF, l'une des entrées essentielles pour l'application du framework est la liste de recommandations $R(u) = n, s(u, n)$ comprenant les scores de pertinence de chaque news non lue pour chaque utilisateur. Je rappelle également que le framework ADF est agnostique à l'algorithme de recommandation appliqué, et donc que tout algorithme peut être utilisé pour le processus de recommandation en amont de l'application d'ADF. Pour la validation expérimentale, j'ai fait le choix d'adopter une approche basée sur le contenu, qui est l'approche de recommandation la plus largement utilisée dans le domaine de la recommandation de news [Raza and Ding, 2022]. Pour cela, j'utilise la bibliothèque ClayRS [19] [Lops *et al.*, 2023], récemment publiée et spécialisée dans la recommandation basée contenu. Parmi les différents algorithmes disponibles, je choisis l'algorithme *CentroidVector*, qui fournit des performances satisfaisantes, est facilement interprétable, extensible à de grands ensembles de données et peu coûteux computationnellement [Zahra *et al.*, 2015]. Ces propriétés font de cet algorithme un bon candidat pour évaluer l'impact du framework ADF sur les recommandations. Concrètement, il calcule la distance entre chaque news n et celles avec lesquelles l'utilisateur u a déjà interagi. La distance entre les news est évaluée à l'aide de la similarité cosinus, et chaque score résultant correspond à un score de pertinence $s(u, n)$, utilisé dans ADF. J'attire l'attention sur la simplicité de l'algorithme sélectionné, contrairement aux algorithmes les plus actuels, souvent basés sur des approches profondes. Cette décision éclairée vise à garantir que l'évaluation ne soit pas perturbée par la complexité de l'algorithme de recommandation sous-jacent. Il me semble qu'opter pour cet algorithme *CentroidVector* est adapté pour la preuve de concept de l'optimisation multi-objectifs proposée dans le framework ADF.

De plus, l'algorithme sélectionné exploitant la représentation vectorielle de chaque news, les embeddings LDA construits à partir des données MIND sont utilisés (voir Annexe C). Pour le processus de recommandation, l'ensemble de données est divisé temporellement pour chaque utilisateur, avec 75% des interactions utilisées pour l'ensemble d'apprentissage et les 25% restants pour l'ensemble de test. Pour les expériences, le score de pertinence $s(u, n)$ est calculé pour chaque news candidate, donc $R = |N|$.

Définition des aspects

Une seconde entrée essentielle à l'application du framework ADF est l'ensemble des aspects A, permettant de caractériser les items et sur lesquels le profil de chaque utilisateur u est construit. Travaillant sur un ensemble de données de news, j'ai fait le choix de définir l'ensemble des aspects A selon les sujets d'actualités traités dans chacune des news (*topics*). Ainsi, dans la suite de ce manuscrit, les aspects seront appelés sujets.

Les actualités du jeu de données MIND ont été catégorisées et sous-catégorisées manuellement, de façon binaire. Cependant, certaines catégories sont peu compréhensibles et sous-représentées dans le jeu de données. Ainsi, afin d'affiner l'identification des sujets, et notamment d'évaluer le degré d'appartenance d'une news à un sujet spécifique, je propose de redéfinir les sujets à l'aide d'une modélisation thématique automatique reposant sur l'exploitation des *embeddings* LDA.

Cette modélisation automatique se déroule en 2 étapes clés :

1. Un algorithme de réduction de la dimension est tout d'abord appliqué sur les *embeddings* LDA : l'algorithme UMAP [20] [McInnes *et al.*, 2018]. Ce dernier est couramment appliqué comme étape de pré-traitement pour le clustering lors de la modélisation thématique. Il

19. https ://swapuniba.github.io/ClayRS/
20. https://umap-learn.readthedocs.io/en/latest/index.html

permet de réduire le nombre de dimensions permettant de caractériser chacune des news, tout en préservant la structure locale et globale des données.

2. Un algorithme de clustering est ensuite appliqué sur les *embeddings* réduits : l'algorithme HDBSCAN [21] [Campello *et al.*, 2013]. Cet algorithme basé sur la densité permet d'identifier certains clusters de faible densité, tout en ignorant les potentielles données bruyantes (*outliers*). De plus, il n'est pas nécessaire de spécifier à l'avance le nombre de clusters attendus, ce qui présente un avantage lorsque la structure thématique des données n'est pas connue en amont.

Ces deux algorithmes (UMAP et HDBSCAN) sont complémentaires et généralement appliqués ensemble pour la modélisation thématique, y compris dans les modèles de langues les plus répandus, comme BERTopic [22].

Les paramètres de ces deux algorithmes (UMAP et HDBSCAN) sont optimisés pour obtenir des performances de clustering optimales, évaluées par l'indice de Silhouette [Rousseeuw, 1987], précédemment utilisé pour la modélisation (Partie I). Cependant, le second indice utilisé précédemment, l'indice de Davies-Boulding n'est pas adapté pour les algorithmes de clustering basés sur la densité comme HDBSCAN. Ainsi, cet indice est remplacé par l'indice DBCV (*Density-Based Clustering Validation*) [Moulavi *et al.*, 2014], qui a été spécialement défini pour évaluer les résultats des algorithmes de clustering basés sur la densité. Une description détaillée de cette phase d'optimisation est donnée dans l'Annexe D.

Dans mon cas, les performances optimales sont obtenues lorsque le nombre de clusters identifiés est de 13 (indice de Silhouette $= 0,23$, indice DBCV $= 0,60$), correspondant à 13 sujets identifiés parmi les news du jeu de données étudié. L'ensemble des aspects A utilisés par le framework ADF est donc l'ensemble des sujets d'actualité, avec $|A| = 13$, où chaque aspect $a \in A$ est interprétable et correspond à un sujet spécifique.

Définition du profil utilisateur et évaluation de la diversité

A partir des aspects précédemment définis et des données d'interactions, qui constituent la troisième entrée essentielle au framework ADF, il est alors possible de construire le profil utilisateur, *c.-à-d.* la *distribution de sélection* des intérêts de u sur l'ensemble des sujets A (Étape 1, Section 5.2.1). Conformément à la proposition de [Eskandanian *et al.*, 2017], je choisis d'évaluer l'intérêt de u pour un sujet a comme le rapport entre l'intérêt de u pour ce sujet et l'intérêt total de u sur A, comme présenté dans l'Équation (5.5).

$$p(a|u) = \frac{\sum_{n \in N_u} s(u,n) q(n|a)}{\sum_{a \in A} \sum_{n \in N_u} s(u,n) q(n|a)} \tag{5.5}$$

où $s(u,n)$ est le score de pertinence fourni en entrée de ADF, N_u est l'ensemble des news avec lesquelles u a interagi. La probabilité $q(n|a)$ est la force d'appartenance de la news n à un sujet a, obtenue lors de l'application de l'algorithme HDBSCAN. Cette probabilité permet d'évaluer plus finement l'intérêt de u pour un sujet spécifique. Je rappelle ici que le profil de u (*distribution de sélection*) est $P(u) = \{p(a|u)\ \forall a \in A\}$.

Une fois ce profil utilisateur construit, la seconde étape consiste à calculer la diversité de sélection associée (Étape 2, Section 5.2.2). Au vu de la formation du profil utilisateur $P(u)$ sous forme de distribution de probabilité, je fais le choix d'évaluer la diversité de sélection de u à

21. https://hdbscan.readthedocs.io/en/latest/index.html
22. https://maartengr.github.io/BERTopic/index.html

l'aide de l'entropie de Shannon normalisée (Équation (2.1), page 31), comme proposé dans la métrique GRAIL (Chapitre 2). L'incertitude qui est évaluée au travers de la mesure d'entropie permet de quantifier la dispersion de l'intérêt de l'utilisateur sur les différents sujets $a \in A$, et donc d'évaluer la diversité associée. La formule de la diversité basée sur la mesure d'entropie appliquée au profil utilisateur $P(u)$ est présentée dans l'Équation (5.6).

$$diversity(P(u)) = \frac{-\sum_a p(a|u) log_2(p(a|u))}{log(|A|)} \tag{5.6}$$

La valeur de diversité résultante est comprise dans $[0; 1]$. Les valeurs élevées indiquent que u est intéressé de manière homogène par tous les aspects $a \in A$, tandis que les valeurs faibles indiquent que u a des intérêts hétérogènes, avec des intérêts majoritaires pour des aspects spécifiques.

Ces deux premières étapes du framework ADF sont donc appliquées pour l'ensemble des utilisateurs, et restent inchangées pour l'ensemble de l'évaluation. Les étapes 3 à 5 peuvent ensuite être appliquées.

Métriques d'évaluation

Une fois le framework ADF appliqué, il convient évidemment d'évaluer les recommandations qui ont été produites. Ainsi, la taille de la liste de recommandations est fixée à $k = 20$. Cette valeur de k a notamment été sélectionnée pour avoir $k \geq |A| = 13$, ce qui permet de représenter chaque sujet d'actualité, si nécessaire, dans la liste de recommandations finale ($R^*(u)$). De plus, une taille de liste de 20 news me paraît être une taille raisonnable pour la recommandation de news. Il est également important de noter ici que cette valeur peut avoir un impact sur les mesures d'évaluation puisque k peut être plus élevé que le nombre de news accédées dans l'ensemble de test pour certains utilisateurs. Néanmoins, cet impact sera similaire d'un modèle à l'autre, ce qui ne biaise pas l'analyse.

Pour évaluer les performances des recommandations, les trois dimensions clés du framework ADF sont évaluées : exactitude, diversité et équité. Premièrement, pour évaluer l'exactitude, la mesure très répandue de précision, notée *Précision@20* (Équation (4.3), page 91), est appliquée. Cette dernière évalue la capacité du système de recommandation de news à sélectionner des informations pertinentes pour un utilisateur u.

Deuxièmement, pour évaluer la diversité, la métrique très répandue de *diversité intra-liste (ILD)* [Ziegler *et al.*, 2005] (Équation (4.10), page 94), est calculée. Cette métrique mesure la distance entre toutes les paires de news dans la liste de recommandations. Dans le cadre de l'évaluation d'ADF, cette distance est calculée avec la similarité cosinus entre les vecteurs LDA représentant chaque news. Les valeurs *ILD* obtenues informent quant à la **diversité de contenu** des news recommandées. En complément de *ILD*, je propose également d'évaluer la **diversité des sujets** (aspects) recommandés à l'aide de la métrique S-Recall, notée *S-Recall@20* [Zhai *et al.*, 2015] (Équation (4.11), page 94). Cette métrique calcule le ratio des sujets couverts dans la liste de recommandations.

Troisièmement, pour évaluer la dimension d'équité, et suivant la définition sur laquelle repose le framework ADF, je propose d'évaluer dans quelle mesure la distribution des sujets dans la liste de recommandations ($P(R^*_k(u))$) est compatible avec la *distribution de sélection* lissée propre à chaque utilisateur. Pour cela, la métrique appliquée est inspirée de la métrique de calibration proposée par [Steck, 2018], reposant sur la distance de Hellinger (Équation (4.13), page 95). Je rappelle que cette mesure de calibration permet de comparer la distribution des intérêts d'un utilisateur sur les différents aspects, à la distribution des recommandations sur

ces mêmes aspects. Une calibration de 0 indique des distributions semblables, donc une parfaite calibration, traduisant une équité optimale. Dans le cadre de l'évaluation d'ADF, cette calibration est calculée entre la *distribution cible* $P^*(u)$, et la distribution des aspects dans la liste de recommandations $P(R_k^*(u))$.

Lors de l'application de cette métrique de calibration sur les baselines, auxquelles est comparé le framework ADF, la distribution des aspects dans les recommandations n'est pas contrôlé car la contrainte d'équité, au cœur d'ADF n'est pas imposée. Ainsi, pour fournir une évaluation aussi juste que possible, la *distribution de sélection* lissée la plus proche de la distribution des aspects dans les recommandations est identifiée pour les modèles de référence. Pour cela, le paramètre δ est optimisé tel que présenté dans l'Équation (5.7).

$$\delta' = \underset{\delta}{\operatorname{argmin}} \quad C_H(P(R_k^*(u)), smooth(P(u), \delta)) \tag{5.7}$$

La valeur de calibration C_H résultante indique dans quelle mesure la distribution de la liste de recommandations est équitable par rapport à une distribution lissée, compatible avec les intérêts de l'utilisateur.

5.4.3 Phase 1 : comparaison du framework ADF aux baselines

Les résultats sont présentés graphiquement dans les Figures 5.8, 5.9 et 5.11. Je précise que les figures ont été construites de façon à faciliter l'analyse des variations de chaque métrique, indépendamment des autres. Les échelles des figures peuvent donc être très différentes d'une métrique d'évaluation à une autre. Par ailleurs, les résultats détaillés avec les valeurs chiffrées sont donnés dans l'Annexe E.

Concernant l'analyse statistique des résultats, le test de Shapiro-Wilk a tout d'abord été appliqué afin d'évaluer la normalité de la distribution des données. Ce test a permis de conclure que les données ne suivent pas une distribution normale ($p_{valeur} < 0,01$), donc un test non paramétrique est privilégié pour la suite de l'analyse statistique. Les comparaisons étant effectuées sur un même ensemble d'utilisateurs (échantillons appariés), c'est le test de Wilcoxon qui est appliqué, et le seuil de signification est fixé à $0,01$. Lors de la présentation des résultats, lorsque ces derniers sont dits significatifs, cela signifie que $p_{valeur} < 0,01$.

A-baseline *vs.* AF-baseline

Cette première phase de validation a pour objectif d'évaluer l'impact de l'optimisation de l'équité seule sur les performances des recommandations, par rapport à une approche commune de maximisation de l'exactitude. La A-baseline, comme expliqué, s'appuie uniquement sur les scores de pertinence $s(u,n)$ calculés avec l'algorithme de recommandation *CentroidVector* et vise à maximiser l'exactitude. La valeur de précision obtenue est de $Précision@20 = 0,224$ (Figure 5.8a).

Par ailleurs, avec une liste de recommandations entièrement calibrée (AF-baseline), *c.-à-d.* dont la distribution est égale à la distribution de sélection de u ($\alpha = 0$ dans ADF), la précision est significativement augmentée de 1% avec $Précision@20 = 0,227$ (Figure 5.8a). Certaines news pour lesquelles les scores de pertinence sont inférieurs au $top - 20$ sont donc plus pertinentes pour les utilisateurs, et ne sont recommandées qu'en optimisant la dimension d'équité. De plus, la diversité *ILD* de la AF-baseline est significativement plus élevée de 25% par rapport à celle de la A-baseline ($0,450$ *vs.* $0,562$), et la diversité *S-Recall@20* est significativement améliorée de 37% ($0,383$ *vs.* $0,525$)(Figures 5.8b et 5.8c). L'optimisation de l'équité permet donc d'apporter

de la diversité dans les recommandations, sans que cette dernière ne soit explicitement optimisée. Finalement, en ce qui concerne l'équité, la valeur de calibration C_H est significativement plus faible pour la AF-baseline avec $C_H = 0,141$, contre $C_H = 0,364$ avec la A-baseline (Figure 5.8d). Je rappelle ici qu'une valeur de C_H plus faible traduit une équité plus élevée. Cette faible valeur de C_H est attendue pour la AF-baseline, puisque la distribution des sujets d'actualité dans la liste de recommandations est contrainte par la *distribution de sélection* de u, et elle est donc équitable. Je précise également que pour cette AF-baseline, C_H est différent de 0, mais reste proche de 0 car lors de l'étape de ré-ordonnancement, l_a est arrondi à l'entier le plus proche et la distribution exacte peut ne pas être respectée (Section 5.2.5).

Cette première étape de l'évaluation permet de conclure qu'**une calibration complète ($\alpha = 1$ avec ADF), en plus d'assurer la dimension d'équité, contribue à une augmentation significative de la diversité, à la fois du contenu et des sujets recommandés, et avec même une augmentation inattendue de l'exactitude**.

A-baseline *vs.* AD-baseline

Cette seconde étape de validation a pour objectif d'évaluer l'impact du compromis exactitude/diversité sur les performances des recommandations, par rapport à une approche commune de maximisation de l'exactitude. En effet, la AD-baseline se concentre à la fois sur les dimensions de précision et de diversité, par le biais d'un compromis représenté par λ (Équation (4.14)) : plus λ est élevé, plus les recommandations sont diversifiées. La Figure 5.8a montre que jusqu'à $\lambda = 0,5$, la AD-baseline permet de conserver une précision satisfaisante, proche de celle obtenue avec la A-baseline (0,224), sans perte significative. Une légère augmentation (jusqu'à $Précision@20 = 0,225$) est même atteinte, à la fois pour $\lambda = 0,3$ et $\lambda = 0,4$ (Figure 5.8a). Cependant, dès lors que $\lambda > 0,5$, la précision chute significativement jusqu'à 33%, avec $Précision@20 = 0,173$ lorsque $\lambda = 1$.

Suivant ces résultats, je propose de me concentrer sur les résultats obtenus avec $\lambda = 0,5$, la valeur λ la plus élevée, donc permettant l'apport en diversité le plus important, qui n'est pas associée à une diminution de la précision. Avec cette valeur de λ, la diversité de contenu est significativement plus élevée de 5% par rapport à la A-baseline ($ILD = 0,472$ *vs.* $ILD = 0,450$) (Figure 5.8b). Cette augmentation de la diversité de contenu va de pair avec une augmentation significative de 3% de la diversité des sujets, puisque *S-Recall@20* passe de 0,383 avec la A-baseline à 0,395 avec $\lambda = 0,5$ (Figure 5.8c). La liste de recommandations ré-ordonnée avec l'approche *greedy* est donc plus diversifiée à la fois en termes de contenus et de sujets. En ce qui concerne l'équité, la métrique de calibration C_H est significativement plus basse pour la AD-baseline (0,349) que pour la A-baseline (0,364) (diminution de 3%) lorsque $\lambda = 0,5$. Je rappelle ici que l'équité n'est optimisée par aucune de ces deux baselines, ce qui explique que ces valeurs de C_H soient relativement élevées. Néanmoins, l'apport en diversité dans les recommandations semble participer à une amélioration de l'équité. Ces résultats sont cohérents avec ceux présentés par [Kaya and Bridge, 2019a].

Pour conclure cette étape de validation, je propose d'étudier les performances lorsque l'apport en diversité est élevé. Avec $\lambda = 0,9$, le gain en diversité est important et significatif : 36% d'augmentation pour la métrique ILD (0,614 *vs.* 0,450), et 28% pour la métrique *S-Recall@20* (0,492 *vs.* 0,383). Ce fort apport en diversité est pourtant associé à un faible impact sur la précision : une perte de seulement 3% de la valeur de $Précision@20$ est observée (0,217 *vs.* 0,224) (Figure 5.8a). Finalement, avec cette valeur de λ, la liste de recommandations est également significativement plus équitable de 13 % (0,316 *vs.* 0,364). Cependant, lorsque l'apport en diver-

sité est maximal, avec $\lambda = 1$, les diversités à la fois de contenus et de sujets sont, comme attendu, largement améliorées, mais l'impact sur la précision est fort, avec une baisse significative de 23% ($0,173$ vs. $0,224$), associée à des recommandations moins équitables ($C_H = 0,370$).

En conclusion, avec $\lambda = 0,5$, le compromis exactitude/diversité appliqué par l'intermédiaire du ré-ordonnancement *greedy* permet de maintenir une précision satisfaisante et égale à celle de la A-baseline, tout en augmentant la diversité de contenu de 5%, la diversité de sujets de 3%, et en améliorant l'équité de 3%. Cependant, des valeurs de λ plus élevées améliorent encore la diversité et l'équité, mais au fort détriment de la précision. Lorsque l'apport en diversité est maximal et que la précision n'est plus prise en compte ($\lambda = 1$), le fort impact négatif sur la précision s'accompagne d'une baisse de l'équité. **L'application d'un compromis exactitude/diversité est donc possible et permet d'apporter de la diversité en conservant une exactitude acceptable, tout en améliorant l'équité des recommandations, qui n'est pourtant pas explicitement optimisée.**

AF-baseline *vs.* AD-baseline

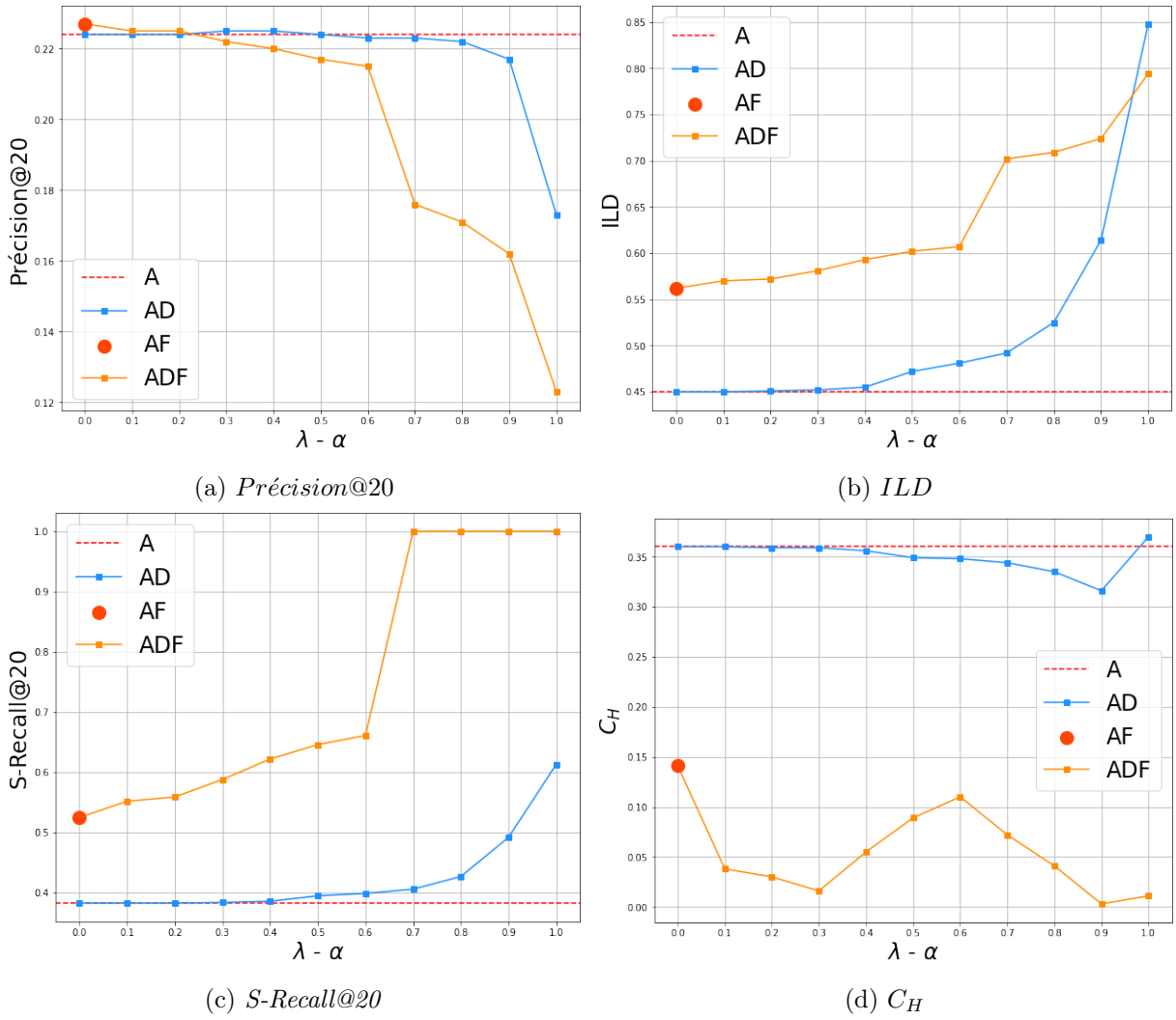
Cette troisième étape de validation a pour objectif de comparer les performances obtenues lorsque l'équité ou la diversité est optimisée. La comparaison de la AF-baseline et la AD-baseline, permet ainsi d'identifier que la différence de précision est limitée jusqu'à $\lambda = 0,7$ ($0,227$ avec la AF-baseline, et $0,223$ avec la AD-baseline (Figure 5.8a). Avec cette valeur de $\lambda = 0,7$, la diversité du contenu (*ILD*) est 14% significativement plus élevée ($0,492$ *vs.* $0,562$) (Figure 5.8b), et la diversité des sujets (*S-Recall@20*) est 29% significativement plus élevée ($0,406$ *vs.* $0,525$) avec la AF-baseline (Figure 5.8c). Ces résultats sont quelque peu inattendus puisque cette approche n'optimise pas la diversité, mais uniquement l'équité. L'optimisation de l'équité permet également, comme attendu, de garantir des recommandations plus équitables ($C_H = 0,141$ avec la AF-baseline *vs.* $C_H = 0,344$ avec AD-baseline et $\lambda = 0,7$) (Figure 5.8d), tout en participant à augmenter la diversité.

Ces résultats mettent à nouveau en avant la **corrélation naturelle entre la diversité et l'équité.** Pour faire suite à cette conclusion préliminaire, et pour évaluer la qualité de l'approche de diversification proposée dans le framework ADF, je propose donc d'évaluer si la diversité sous contrainte d'équité permet d'apporter davantage de diversité, et quel en serait l'impact sur l'exactitude de ces recommandations.

ADF *vs.* A-baseline, AF-baseline et AD-baseline

Pour compléter l'évaluation d'ADF, je propose dans cette quatrième étape de validation d'analyser l'impact de l'apport en diversité contraint par l'équité sur les performances des recommandations. Pour cela, les performances obtenues avec les différents valeurs de $\alpha \in [0;1]$ ($\beta = 0$) sont analysées puis comparées aux performances obtenues avec les trois baselines. Je rappelle ici que l'apport en diversité proposé par ADF est au niveau des aspects, donc des sujets des news dans le contexte d'application.

Pour commencer, je propose de comparer les performances obtenues lorsque la calibration est complète ($\alpha = 0$) et lorsque l'apport en diversité est maximal ($\alpha = 1$). Ainsi, je rappelle que lorsque $\alpha = 0$, correspondant à la AF-baseline, il en résulte une augmentation significative des trois dimensions optimisées par rapport à la A-baseline : la précision augmente d'1%, la diversité de contenu augmente de 25%, la diversité des sujets augmente de 37%, et l'équité est améliorée



(a) *Précision*@20 (b) *ILD*

(c) *S-Recall@20* (d) C_H

FIGURE 5.8 – Valeurs des métriques pour la A-baseline, la AD-baseline, la AF-baseline, et ADF.

de 61%. À l'inverse, lorsque α=1, l'impact sur la précision est très important : la valeur de *Précision*@20 chute de 45%, passant de $0,224$ à $0,123$ (Figure 5.8a). Cette baisse importante de la précision est attendue pour un apport en diversité très élevé, et suit une pente similaire à celle de la perte de précision obtenue lors de l'application de AD-baseline (Figure 5.8a). En ce qui concerne l'apport en diversité, la diversité de contenu augmente de 76%, les valeurs *ILD* passant de $0,450$ à $0,794$ (Figure 5.8b). Par ailleurs, la diversité de sujets est maximale, *S-Recall@20*$= 1$, car tous les aspects $a \in A$ sont représentés dans la liste de recommandations diversifiées (Figure 5.8c). Enfin, les valeurs de calibration C_H diminuent de 46%, passant de $0,364$ à $0,011$ (Figure 5.8d). Ceci indique que les distributions dans les listes de recommandations, $P(R_k^*(u))$ restent équitables selon notre définition de l'équité (Section 5.2.4), même si ces listes de recommandations sont très diversifiées.

Maintenant que les résultats obtenus avec les deux valeurs extrêmes de α sont analysés, je propose de me focaliser sur les valeurs intermédiaires de ce paramètre. Jusqu'à $\alpha = 0,2$, la précision n'est pas diminuée, elle est même significativement plus élevée de $0,4\%$ par rapport à la AD-baseline (*Précision*@20 passant de $0,224$ à $0,225$). Cette amélioration de l'exactitude est

associée à une augmentation de la diversité : les valeurs *ILD* augmentent de 27%, passant de $0,450$ à $0,572$, et les valeurs de *S-Recall@20* augmentent également de 46%, passant de $0,383$ à $0,559$. Cette augmentation plus importante de *S-Recall@20* est attendue puisque la diversification proposée par ADF repose sur les aspects, qui sont ici les sujets des news recommandées. Enfin, avec cette valeur de $\alpha = 0,2$, la métrique de calibration C_H est égale à $0,030$ (Figure 5.8d), confirmant que la liste de recommandations est équitable.

Je compare désormais ces résultats à ceux obtenus avec la AD-baseline lorsque $\lambda = 0,4$. Cette valeur de λ est sélectionnée car elle permet d'avoir une valeur de précision identique ($Précision@20 = 0,225$). L'application du framework ADF permet une augmentation significative de la diversité de contenu de 26% par rapport à la AD-baseline ($0,450$ *vs.* $0,572$), et à une augmentation significative de 45% de la diversité des sujets ($0,386$ *vs.* $0,559$). Ainsi, bien que la nature de la diversité apportée par l'intermédiaire des deux approches soit différente (diversité de contenu pour la AD-baseline, et diversité de sujets pour ADF), l'application du framework ADF permet d'augmenter les deux diversités, tout en assurant l'équité des recommandations.

Par ailleurs, lorsque α augmente et jusqu'à $\alpha = 0,6$, la baisse des valeurs de $Précision@20$ est limitée à 4% ($Précision@20 = 0,215$). Par rapport à la AD-baseline, lorsque $\alpha <= 0,6$ et $\lambda <= 0,6$, les valeurs de $Précision@20$ sont proches (Figure 5.8a). Cependant, la diversité des contenus et des sujets augmente davantage avec ADF. Par exemple, lorsque $\lambda = \alpha = 0,6$, $ILD = 0,481$ avec la AD-baseline, tandis que $ILD = 0,607$ avec ADF, ce qui représente une augmentation significative de 26%. De la même façon, *S-Recall@20*$= 0,661$ avec ADF, contre *S-Recall@20*$= 0,399$ avec la AD-baseline, ce qui représente une augmentation significative de 66%. En ce qui concerne la dimension de l'équité, comme prévu, ADF permet de maintenir l'équité au travers de la contrainte imposée, et les valeurs de calibration C_H restent faibles pour l'ensemble des valeurs de α. Ces résultats ne sont pas retrouvés avec la AD-baseline, pour laquelle les valeurs de C_H restent élevées ($> 0,3$), quelle que soit la valeur de λ.

Finalement, tout comme pour la AD-baseline, lorsque l'apport en diversité est trop élevé ($\alpha >= 0,6$), la précision est considérablement réduite. Par exemple, avec $\alpha = 0,9$, $Précision@20 = 0,162$, ce qui représente une baisse significative de 28%.

En résumé, **l'application du framework ADF permet de fournir des recommandations significativement plus diversifiées, à la fois en termes de contenus et de sujets, tout en étant équitables et sans impact sur l'exactitude de ces recommandations.** Ainsi, l'approche proposée consistant à contraindre la diversité par l'équité n'a pas nécessairement un impact négatif sur l'exactitude, tant que l'apport en diversité est modéré. Ces résultats, en plus de mettre en avant la qualité du framework ADF, permettent également de confirmer que le caractère myopique de l'approche *greedy* ne permet pas de gérer la nature de la diversité, limitant la qualité du compromis entre exactitude et diversité [Seymen *et al.*, 2021]. Au contraire, l'approche par contrainte proposée dans ADF permet de contrôler l'apport en diversité tout en assurant le respect des intérêts de l'utilisateur au travers de l'équité.

5.4.4 Phase 2 : comparaison d'une approche personnalisée de la diversité à une approche globale de diversité

Pour compléter l'évaluation précédente, cette seconde phase a pour objectif d'évaluer les avantages de la diversification personnalisée, par rapport à une diversification globale (lorsque la valeur de *diversité cible* est la même pour tous les utilisateurs). Pour rappel, les diversités globales étudiées sont comprises entre $0,6$ et 1, avec $\Delta = 0,1$.

Concrètement, les performances obtenues lors de la définition d'une *diversité cible* globale

sont comparées à celles obtenues avec la configuration d'ADF (valeur de α) permettant d'avoir la valeur de *S-Recall@20* la plus proche. Les deux approches sont comparables dans le sens où elles offrent la même diversité en moyenne.

En ce qui concerne la dimension d'équité, je ne la discuterai pas lors de la présentation des résultats puisque les valeurs de calibration restent faibles pour un apport en diversité global lorsque l'approche de diversification contrainte par l'équité, proposée dans ADF, est appliquée.

Diversité globale *vs.* diversité personnalisée

Pour une diversité cible globale fixée à $0,6$, il n'est pas pertinent de comparer les performances avec celles obtenues avec ADF puisque cette diversité cible est inférieure à la moyenne des *diversités de sélection* dans le jeu de données étudié, $\overline{diversity(P(u))} = 0,653$. Cela ne correspond pas aux caractéristiques de la fonction $f()$, énumérées dans la section 5.2.3.

Pour une *diversité cible* globale de $0,7$, la valeur de *S-Recall@20* est de $0,558$, qui est proche de la valeur obtenue avec $\alpha = 0,2$ lors de l'application d'ADF (Figure 5.9c). Pour une diversité de sujets similaires, la diversité de contenu est significativement moins élevée de 1% avec ADF ($ILD = 0,572$ *vs.* $ILD = 0,577$) (Figure 5.9b). Par ailleurs, la *Précision*@20 obtenue avec ADF est significativement plus élevée de 1% ($0,225$ *vs.* $0,223$) (Figure 5.9a). Ce bénéfice limité de l'approche personnalisée peut s'expliquer par la valeur de diversité cible ($0,7$) proche de la diversité de sélection moyenne ($0,653$). Je m'intéresse donc à l'impact lorsque la diversité cible est plus élevée.

Lorsque la diversité cible globale est plus élevée et fixée à $0,8$, la valeur de *S-Recall@20*$= 0,660$ est similaire à celle obtenue avec $\alpha = 0,6$ (Figure 5.9c). De la même façon que précédemment, la diversité de contenu est inférieure avec l'approche ADF ($ILD = 0,607$ *vs.* $ILD = 0,615$) (Figure 5.9b), mais la précision avec l'approche ADF est supérieure de 3% ($Précision@20 = 0,215$ vs. $Précision@20 = 0,211$) (Figure 5.9a).

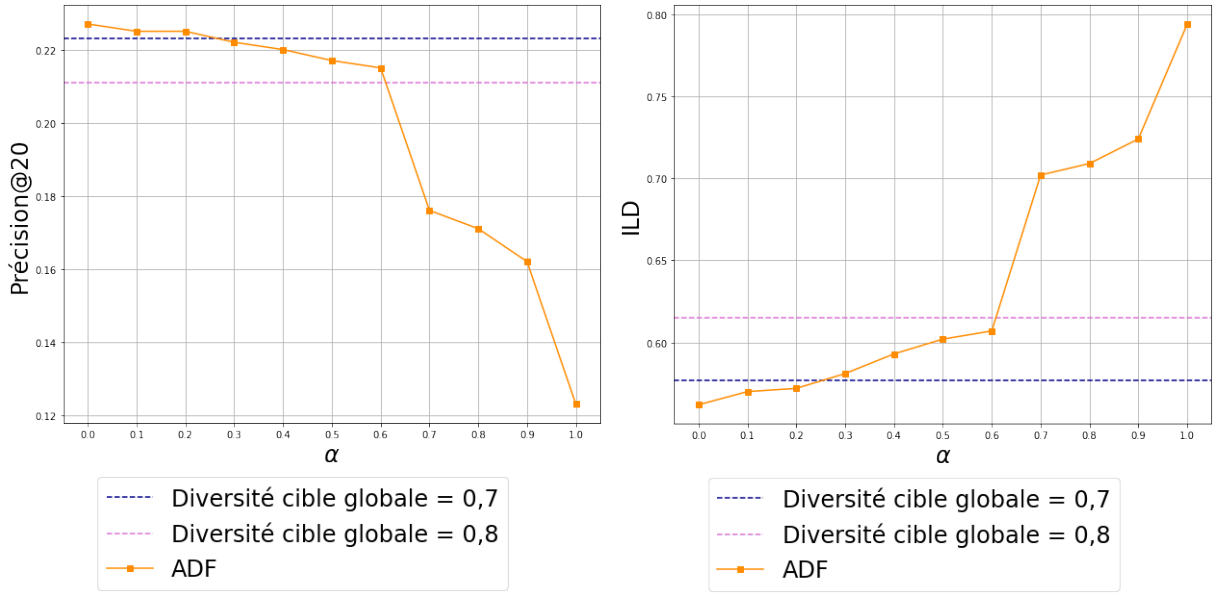
Enfin, pour des valeurs de *diversité cible* très élevées, fixées à $0,9$ et 1, les différences sur les dimensions d'exactitude et de diversité, à la fois de contenu et de sujets, sont limitées entre les deux approches. Cela indique que, lorsque la diversification est très élevée, la baisse de précision est très importante, même si la diversité est personnalisée.

Pour résumer, **bien que l'approche globale de diversité permette un apport légèrement supérieur en diversité par rapport à l'approche personnalisée, elle a un impact significatif sur l'exactitude moyenne des recommandations**. La personnalisation de l'apport en diversité semble donc améliorer l'exactitude des recommandations. Une évaluation plus fine de cet impact sur l'exactitude des recommandations semble cependant nécessaire pour distinguer les potentielles différences de performances en fonction des utilisateurs.

Impact de la personnalisation en fonction de la *diversité de sélection*

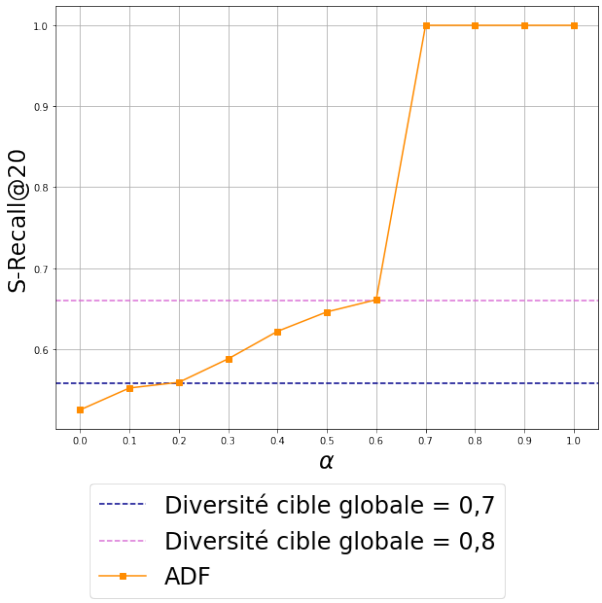
L'objectif de cette dernière étape de validation est d'évaluer si les performances sont similaires pour l'ensemble des utilisateurs avec les deux approches, ou si de fortes divergences peuvent être observées entre les utilisateurs en fonction de leur *diversité de sélection*. Pour répondre à cet objectif, je propose de me concentrer sur les résultats obtenus avec une *diversité cible* globale de $0,8$, qui est celle ayant eu le plus fort impact sur les valeurs de *Précision*@20.

Je compare tout d'abord la distribution des valeurs de *diversité de sélection* avec celles des valeurs de *diversité cible* personnalisée, lorsque $\alpha = 0,6$, qui permet d'avoir une diversité des sujets (*S-Recall@20*) similaire à celle obtenue lorsque la diversité cible globale est fixée à $0,8$,



(a) *Précision*@20

(b) *ILD*



(c) *S-Recall@20*

FIGURE 5.9 – Valeurs des métriques lorsque la *diversité cible* globale est intermédiaire ($0,7$ et $0,8$) et pour ADF.

avec *S-Recall@20*$= 0,661$. L'étude de ces distributions permet d'observer que, lorsque la valeur de *diversité cible* est personnalisée, les valeurs de diversité se distribuent entre $0,26$ pour les utilisateurs ayant les diversités de sélection les plus faibles, et $0,93$ pour les utilisateurs ayant les diversités de sélection les plus élevées (Figure 5.10). La diversité cible moyenne est de $0,77$.

Ainsi, plus de 60% des utilisateurs ont une valeur de diversité cible personnalisée inférieure à la diversité cible globale fixée à $0,8$. Je m'interroge donc sur l'impact d'une approche non

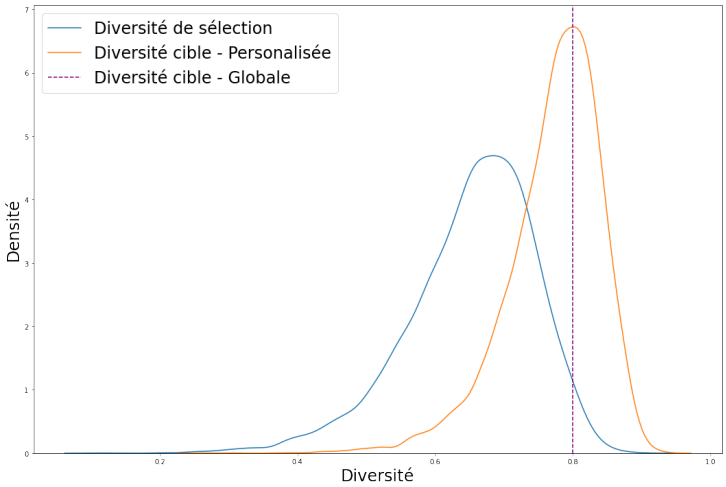


FIGURE 5.10 – Distribution des *diversité de sélection* et *diversité cible*, personnalisée et globale.

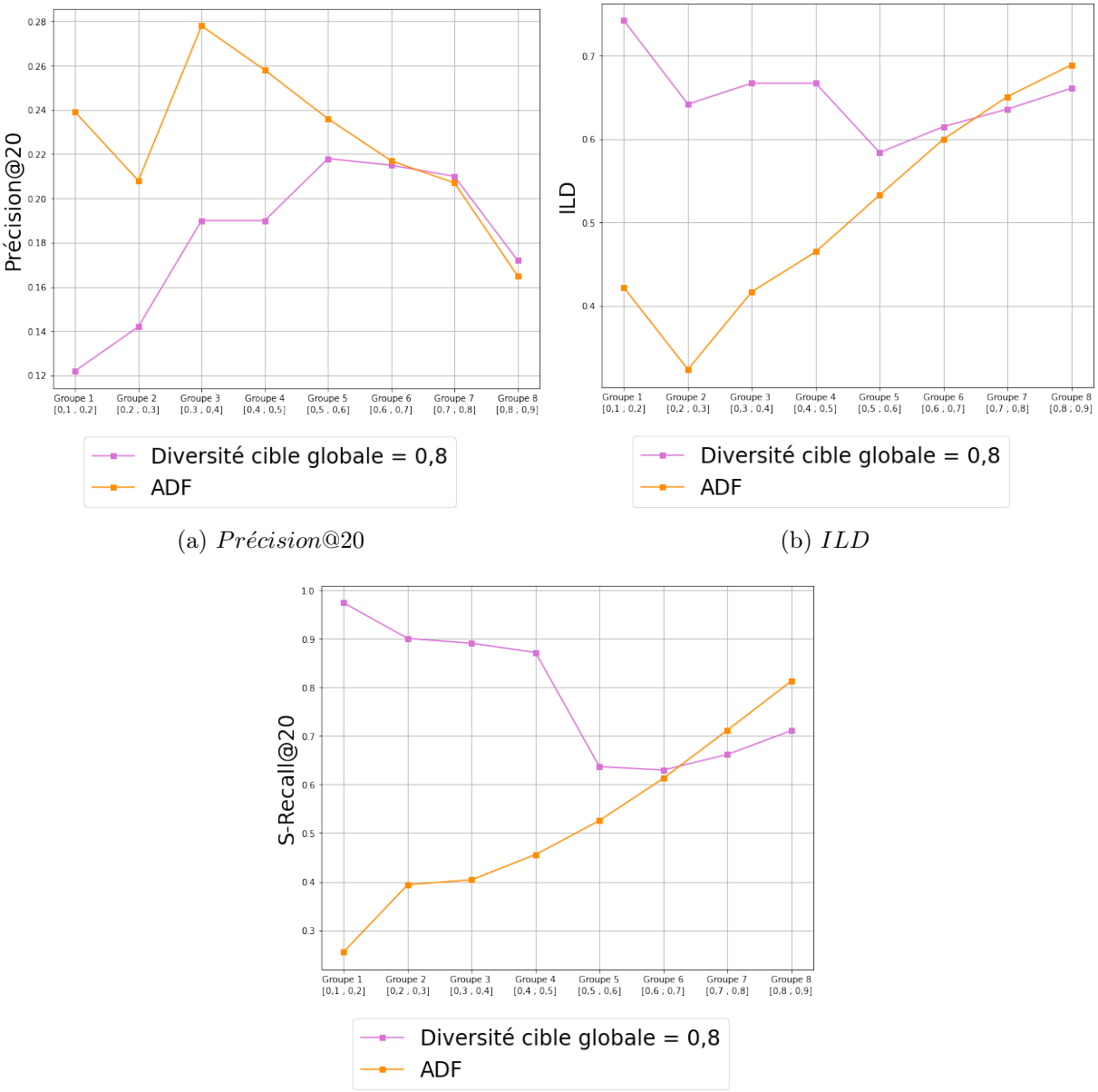
personnalisée sur ces utilisateurs.

Pour répondre à cette interrogation, je divise les utilisateurs du jeu de données en groupes en fonction de leur *diversité de sélection*, puis j'étudie les performances de recommandation pour chacun de ces groupes. Les valeurs de *diversité de sélection* sont comprises entre $0,1$ et $0,9$. Les utilisateurs ont été regroupés selon des intervalles de diversité : $[0,1;0,2[$, $[0,2;0,3[$, *etc.* Au total, ce sont 8 groupes d'utilisateurs de tailles différentes qui sont formés, et pour lesquels les performances sont évaluées, à l'aide des métriques *Précision*@20, ILD et *S-Recall@20*. Ayant expliqué plus tôt dans cette section que l'équité est assurée quelle que soit la nature de la diversité cible, globale ou personnalisée, cette dernière n'est pas évaluée ici. Le détail des groupes et des performances est donné dans l'Annexe E.

Dans un premier temps, l'analyse des performances permet de mettre en évidence des valeurs de *Précision*@20 supérieures avec ADF ($\alpha = 0,6$) par rapport aux performances obtenues lorsque la diversité cible est globale, pour tous les groupes d'utilisateurs dont la *diversité de sélection* est inférieure à $0,7$ (Figure 5.11a).

Dans un second temps, l'évolution des deux métriques de diversité au sein des groupes est progressive pour ADF : plus la diversité de sélection des utilisateurs du groupe est élevée, plus la diversité de contenu (ILD) et la diversité de sujets (*S-Recall@20*) est élevée dans les recommandations (Figures 5.11b et 5.11c, respectivement). Ce caractère progressif de la diversité dans les recommandations en fonction de la diversité de sélection n'est pas retrouvé avec l'approche globale : la diversité, quelle qu'elle soit, n'est pas corrélée avec la *diversité de sélection*. Ces résultats sont attendus puisque dans cette approche globale, l'apport n'est pas personnalisé.

Ainsi, l'absence de personnalisation dans les recommandations entraîne une diminution importante des valeurs de *Précision*@20, probablement liée à un apport trop conséquent en diversité pour les utilisateurs dont la diversité de sélection est inférieure à la diversité cible globale fixée. Ce résultat met ainsi en avant que **la personnalisation de l'apport en diversité permet de garantir l'exactitude des recommandations pour l'ensemble des utilisateurs, au contraire de l'approche non personnalisée, dont l'exactitude est satisfaisante uniquement pour les utilisateurs dont la *diversité de sélection* est proche de la diversité globale ciblée. En plus de garantir une augmentation de la diversité et le maintien de l'équité, le framework ADF permet donc de conserver l'exactitude des recom-**



(a) *Précision*@20

(b) *ILD*

(c) *S-Recall@20*

FIGURE 5.11 – Valeurs des métriques pour les différents groupes d'utilisateurs définis, lorsque la diversité cible globale est à $0,8$ et lorsque $\alpha = 0,6$ avec ADF.

mandations pour l'ensemble des utilisateurs.

5.5 Conclusion et discussion

Dans ce chapitre dédié à la recommandation, un nouveau framework de recommandation, appelé ADF est présenté. Ce framework repose sur une optimisation multi-objectifs de trois dimensions clés de la recommandation : l'exactitude, la diversité et l'équité. La façon dont est construit ce framework permet de répondre à la première sous-question de recherche de ce chapitre

(*QR2.1*).

Comment diversifier les recommandations tout en assurant leur équité ? (*QR2.1*)

L'approche de diversification contrainte par l'équité proposée dans ADF permet d'assurer la compatibilité des recommandations avec les intérêts de l'utilisateur pour certains aspects caractérisant les news. Cette approche inédite par contrainte permet ainsi d'identifier une distribution attendue des news sur les différents aspects dans les recommandations. Cette distribution est établie en fonction d'une diversité cible personnalisée. Les résultats de la validation expérimentale confirment que cette approche permet en effet d'apporter de la diversité, de façon contrôlée, tout en garantissant l'équité.

Au-delà de l'optimisation des dimensions de diversité et d'équité, essentielles à la recommandation de news, l'optimisation de l'exactitude reste primordiale. En effet, lorsque l'exactitude est considérée comme unique dimension, elle participe à l'enfermement des utilisateurs dans des bulles de filtre et donc au phénomène de polarisation. Néanmoins, cette dimension se doit d'être préservée afin de satisfaire les attentes des utilisateurs. La validation expérimentale a ainsi permis de répondre à la seconde sous-question de recherche de ce chapitre (*QR2.2*), directement en lien avec cette dimension d'exactitude.

Quel est l'impact de la diversification équitable sur l'exactitude des recommandations ? (*QR2.2*)

Les résultats ont montré que la diversité apportée par l'application de l'approche ADF permet de conserver une exactitude satisfaisante, parfois même supérieure à celle du modèle baseline ne cherchant qu'à optimiser cette dimension. Le framework ADF répond donc bien à une optimisation multi-objectifs allant au-delà des traditionnels compromis les plus communément appliqués dans la littérature.

La contribution présentée dans ce chapitre répond donc au besoin d'une personnalisation de l'apport en diversité dans les recommandations, et en particulier dans les recommandations de news. Cette personnalisation adresse dans le même temps les aspects éthiques liés à la recommandation, en assurant de ne pas orienter l'intérêt des utilisateurs. Dans la perspective de réduction du phénomène de polarisation ayant guidé l'ensemble de mes travaux de thèse, l'approche de diversification proposée dans ADF permet ainsi d'offrir aux utilisateurs une plus grande diversité de news, tout en préservant leurs intérêts initiaux et en leur fournissant des informations en accord avec leurs attentes. Ces travaux répondent ainsi à ma seconde question de recherche (QR2).

Chapitre 6

Conclusion et perspectives

Sommaire

6.1 Conclusion

Les travaux présentés dans ce manuscrit abordent le phénomène social de la polarisation en ligne, qui est renforcé par l'utilisation de l'IA. C'est notamment l'impact de la personnalisation de l'information présentée, reposant sur des systèmes de recommandation de news, qui est questionné. Bien qu'essentielle, cette personnalisation participe à l'enfermement des utilisateurs au sein de bulles de filtre, où ils n'ont qu'une perception partielle de la réalité, renforçant le phénomène de polarisation. Dans l'objectif de promouvoir un débat plus équilibré et informé, une approche de la littérature consiste à modifier la façon dont est personnalisé le contenu informatif recommandé aux utilisateurs des médias en ligne. Dans cette thèse je défends l'idée que le développement de telles solutions dépolarisantes repose sur une compréhension fine des comportements de polarisation.

Tout au long de ma thèse, et pour l'ensemble des travaux présentés dans ce manuscrit, j'ai eu à cœur d'adopter une approche centrée utilisateur. C'est pourquoi j'ai porté une attention particulière à la modélisation des comportements, me permettant d'en avoir une meilleure compréhension avant de travailler sur l'adaptation des systèmes de recommandation de news afin de réduire le phénomène de polarisation. Ces systèmes sont à forts enjeux de par leur utilisation quotidienne et ubiquitaire et doivent selon moi être conçus en portant une attention particulière à l'éthique pour éviter les biais ou les effets indésirables. Cette ambition est d'ailleurs largement documentée dans le Règlement européen sur l'Intelligene Artificielle (EU AI Act [23]).

23. https://artificialintelligenceact.eu/fr/

Mon travail de thèse a ainsi tout d'abord porté sur la modélisation individuelle, multi-factorielle et temporelle du phénomène de polarisation, pour répondre à ma première question de recherche (*QR1*) :

> ***Question de Recherche 1 (QR1)***
> Comment modéliser les comportements de polarisation individuels de façon multi-factorielle pour tenir compte de la complexité du phénomène de polarisation, et de façon temporelle pour rendre compte de la dynamique sous-jacente ?

Pour y répondre, j'ai tout d'abord proposé une modélisation individuelle et multi-factorielle. Cette modélisation repose sur l'évaluation de facteurs de polarisation au travers d'une mesure d'entropie, permettant d'évaluer la diversité des interactions de chaque utilisateur de façon plus fine. Ces facteurs de polarisation permettent de distinguer différentes classes de comportements, dont la discrimination est améliorée par l'application d'une transformation polynomiale des valeurs d'entropie. Finalement, les différents facteurs de polarisation sont assemblés en une mesure unique de polarisation à l'aide d'un modèle additif généralisé. Il en résulte une métrique individuelle et multi-factorielle de polarisation appelée GRAIL.

L'évaluation expérimentale confirme que les facteurs de polarisation modélisés au travers des composants de la métrique GRAIL permettent de discriminer différentes classes de comportements distinctes. L'identification de ces classes n'étant pas possible à partir des approches de modélisation décrites dans la littérature participe à une compréhension plus fine de la polarisation. Par ailleurs, la capacité à expliquer la variance des valeurs de GRAIL dans chacune des classes identifiées permet à la fois de valider la pertinence de la métrique proposée pour la quantification de la polarisation, mais aussi de fournir une description détaillée des comportements adoptés et des facteurs influençant leur adoption. Le travail présenté dans ce chapitre soumet donc deux contributions, avec d'une part une modélisation multi-factorielle et individuelle des comportements de polarisation, et d'autre part la métrique de polarisation sur laquelle elle repose.

Pour compléter, j'ai proposé une modélisation temporelle du phénomène de polarisation. La polarisation est alors abordée comme un processus évolutif, et les comportements de polarisation identifiés sont comparés au cours de fenêtres temporelles successives. Cette approche temporelle permet notamment de caractériser des périodes de polarisation, définies en fonction du nombre et de la nature des classes de comportement de polarisation identifiées sur une période de temps. Les quatre périodes de polarisation caractérisées (non structurée, équilibrée, de convergence, polarisée) suivent une organisation spécifique. Leur existence et leur durée dépendent de la maturité du débat sur lequel les utilisateurs se polarisent : plus le débat est mature, moins les utilisateurs sont susceptibles d'adopter des positions intermédiaires. Une analyse approfondie de ces périodes a également permis de mettre en évidence l'impact du contexte dans l'évolution des comportements de polarisation en ligne. L'apparition d'événements perturbateurs liés au débat polarisant entraîne en effet une réorganisation des classes de comportements et la transition vers des périodes de polarisation spécifiques. Par ailleurs, l'identification de dynamiques de polarisation à partir de valeurs de polarisation GRAIL permet de quantifier l'évolution temporelle de la polarisation des utilisateurs des réseaux sociaux, qui finissent systématiquement par se polariser au sein d'une communauté spécifique. Ces travaux, constituant la troisième contribution de mon travail de thèse, ont ainsi permis d'affirmer le caractère évolutif de la polarisation, et participent à une modélisation plus fine du phénomène.

Ce travail de modélisation a mis en lumière l'adoption de comportements de polarisation distincts de la part des utilisateurs, susceptibles d'évoluer temporellement. Ces résultats offrent de nouvelles opportunités pour la recommandation de news. En particulier, ils confirment le besoin d'approches personnalisées de recommandation. Pour faire suite à ce travail, je me suis donc intéressée aux approches de recommandation de news. Plus précisément, j'ai travaillé sur l'apport personnalisé en diversité, garantissant le respect de l'équité des recommandations vis-à-vis des intérêts de chaque utilisateur. Ce travail a été guidé par la seconde question de recherche abordée dans ce manuscrit (*QR2*) :

> ***Question de Recherche 2 (QR2)***
> Comment modifier les approches de recommandation de news de façon à apporter de la diversité tout en respectant les préférences des utilisateurs et les aspects éthiques ?

Pour y répondre, j'ai introduit une approche de diversification personnalisée, appelée framework ADF. Ce framework repose sur une optimisation multi-objectif visant à garantir des recommandations qui soient à la fois exactes, précises et équitables. Pour cela, les recommandations calculées par un système de recommandation, quel qu'il soit, sont ré-ordonnées selon une approche de diversification contrainte par l'équité : la diversité est apportée au travers de recommandations dont la distribution sur les différents aspects des news est plus homogène mais compatible avec celle des interactions de l'utilisateur cible. La personnalisation de cet apport en diversité est assurée par une fonction permettant d'estimer une diversité cible pour chaque utilisateur en fonction de sa diversité de sélection initiale. La validation expérimentale confirme que l'application du framework ADF permet d'apporter de la diversité sans impacter l'exactitude des recommandations, et en assurant leur équité. Au contraire des approches traditionnelles avec lesquelles le framework ADF a été comparé, l'apport en diversité est contrôlé et garantit des recommandations compatibles avec les intérêts des utilisateurs. Dans une perspective de dépolarisation, le framework ADF fournit donc des recommandations de news répondant aux attentes des utilisateurs (exactitude), tout en les confrontant à des contenus plus divers (diversité) et sans orienter leurs opinions (équité). Ce framework de recommandation constitue la quatrième contribution de ma thèse.

L'ensemble des travaux présentés participent ainsi à répondre à la problématique de recherche énoncée en introduction :

> **Problématique**
> Comment participer à réduire le phénomène de polarisation en ligne en confrontant les utilisateurs à un contenu plus diversifié ?

En résumé, j'ai montré qu'une réduction du phénomène de polarisation passe tout d'abord par une compréhension fine des comportements sous-jacents, puis par une adaptation personnalisée et éthique des systèmes d'IA qui participent à le renforcer.

L'ensemble des travaux présentés dans ce manuscrit a contribué à renforcer mon attrait pour les problématiques liées au numérique et à l'humain. En ce sens, j'aimerais finalement mettre en avant la richesse que m'ont apportée les différentes collaborations menées, concrétisées par l'adoption d'une approche pluri-disciplinaire. Les questions, hypothèses, méthodologies d'évaluation ou encore analyses que j'ai pu présenter au travers de ce manuscrit n'auraient pas été les mêmes sans les perspectives issues des sciences sociales et sciences politiques.

6.2 Discussion et enjeux scientifiques

J'aimerais discuter ici les enjeux scientifiques soulevés par les contributions de cette thèse.

Pour commencer, la modélisation multi-factorielle et individuelle a été évaluée dans le contexte particulier des réseaux sociaux. Son application à des contextes différents, et notamment en lien avec la recommandation de news, permettrait de compléter son évaluation. Cette adaptation à différents contextes est rendue possible par la généralisabilité de la métrique GRAIL, permise par la nature de ses composants.

Par ailleurs, l'application de l'approche de modélisation individuelle et multi-factorielle présentée sur des débats opposant plus de deux communautés permettrait d'évaluer des contextes politiques plus complexes et au plus proche de la réalité. En effet, dans certains cas et notamment dans le système politique français, l'organisation est complexe et va bien au-delà d'une opposition binaire entre deux partis. Une modélisation des comportements de polarisation dans un tel contexte peut s'avérer délicate de part la pluralité des communautés impliquées, la multiplicité des débats, l'alignement partiel des enjeux, *etc.* La quantification de la polarisation se doit alors de refléter de multiples paramètres pour rester pertinente et permettre la modélisation des comportements de polarisation adoptés.

Les enjeux liés à la modélisation individuelle et multi-factorielle dans des contextes complexes s'appliquent également à la modélisation temporelle. La modélisation temporelle des comportements de polarisation dans des périodes à forts enjeux politiques, comme lors d'élections (présidentielles, européennes, législatives, *etc.*), serait par exemple d'un grand intérêt pour étudier l'évolution des opinions et la manière dont ces dernières guident les choix de vote. Une telle modélisation repose sur des données temporelles correspondant à des périodes plus courtes, et dont les interactions sont fortement impactées par des événements de contextes extérieurs, comme les campagnes politiques des candidats. Par ailleurs, une modélisation basée sur l'exploitation de données couvrant une période plus longue, jusqu'à plusieurs années, permettrait de compléter la compréhension des dynamiques de polarisation. Cela nécessite cependant d'être en mesure de collecter, d'analyser et de stocker de larges séries temporelles, reflétant des interactions influencées par de multiples événements.

Finalement, certains enjeux scientifiques concernent le framework ADF. L'application du framework à d'autres jeux de données spécifiques à d'autres domaines pour lesquels un apport en diversité personnalisé est nécessaire, permettrait d'évaluer dans quelle mesure la diversification contrainte par l'équité peut être appliquée dans des contextes de recommandation variés. Par ailleurs, comme expliqué dans la description du framework, ce dernier est agnostique au système de recommandation appliqué puisqu'il consiste en une approche de ré-ordonnancement. Ainsi, l'application du framework sur des recommandations issues de divers algorithmes permettrait de comparer les performances du framework avec différentes approches, et notamment de comparer les approches de filtrage par contenu aux approches de filtrage collaboratif, qui apportent naturellement plus de diversité. Finalement, dans le contexte spécifique de la recommandation de news, une diversification portant sur des aspects différents (sentiments, sources, *etc.*) permettrait d'identifier de potentiels aspects plus propices à un apport en diversité susceptible de réduire le phénomène de polarisation.

6.3 Perspectives

Au-delà des enjeux scientifiques discutés, les travaux présentés offrent des perspectives de recherche. Dans la suite de cette section, je présente tout d'abord certaines perspectives scientifiques applicables à court terme, puis termine par une perspective plus globale relative à l'impact des systèmes de recommandation de news sur le phénomène de polarisation.

6.3.1 Diversification adaptée aux classes de comportement de polarisation

Le framework ADF présenté apporte une diversité personnalisée dans les recommandations, en assurant leur équité avec les intérêts des utilisateurs. Répondant à un objectif de dépolarisation, il reste à déterminer si l'équilibre offert par cette approche de diversification contribue effectivement à réduire l'enfermement des utilisateurs au sein de bulles de filtre, et ainsi aider à éviter l'adoption de comportements extrêmes de polarisation. Une perspective évidente au travail présenté dans ce manuscrit est alors d'exploiter la métrique GRAIL pour l'adaptation de l'apport en diversité permis par le framework ADF. Bien que cette approche de diversification soit informée par les résultats de la modélisation individuelle et multi-factorielle, une exploitation plus directe des résultats pour la tâche de recommandation apparaît intéressante.

Indépendamment, les contributions de cette thèse permettent d'avoir une meilleure compréhension du phénomène de polarisation et des comportements adoptés (GRAIL), et une approche de diversification multi-objectif (ADF). Il reste à déterminer comment ces deux outils essentiels peuvent être combinés afin de fournir une approche de recommandation robuste garantissant une réduction du phénomène de polarisation. Je précise que l'objectif n'est pas de réduire complètement la polarisation, mais de favoriser des échanges plus apaisés et rétablir un terrain d'entente entre les communautés opposées afin d'éviter les conséquences extrêmes de la polarisation.

Pour répondre à cet enjeu, deux perspectives de recherche sont envisagées. D'une part, la métrique GRAIL pourrait servir de métrique d'évaluation des systèmes de recommandation. Cela permettrait notamment de comparer le degré de polarisation des utilisateurs selon diverses approches de recommandation. Elle constituerait ainsi une nouvelle métrique d'évaluation, particulièrement adaptée à la recommandation de news. Néanmoins, lors d'une évaluation hors ligne, la division du jeu de données en ensemble d'apprentissage et de test limite la possibilité d'observer une réelle modification du degré de polarisation chez les utilisateurs. Ces ensembles de données couvrent des périodes courtes de quelques semaines seulement, et les résultats de la modélisation temporelle ont permis de démontrer qu'une variation du niveau de polarisation, lorsqu'il n'y a pas d'événement de contexte majeur, est progressive. L'application de la métrique GRAIL pour évaluer la capacité de dépolarisation d'une approche de recommandation, telle qu'ADF, nécessite donc une évaluation à plus long terme, via une évaluation en ligne notamment. Dans ce cas de figure, l'évaluation complétée par **la métrique GRAIL permettrait d'évaluer la dimension de polarisation, qu'il est difficile d'évaluer pleinement avec les métriques de la littérature.**

D'autre part, une exploitation de la modélisation individuelle et multi-factorielle permise par la métrique GRAIL permettrait une personnalisation plus fine de l'apport en diversité permise par le framework ADF. Une diversification multi-aspects, où les aspects correspondent aux facteurs de polarisation modélisés, apparaît comme intéressante. La mise en place d'une telle stratégie nécessite d'adapter le framework ADF de façon à répondre à l'optimisation multi-objectif initiale, mais pour différents aspects des news. Par ailleurs, une approche empirique envisagée consiste à adapter l'apport en diversité en fonction des classes de comportement identifiées. Ainsi,

l'appartenance d'un utilisateur à une classe conditionnerait le niveau de diversité apporté à cet utilisateur. Par exemple, les utilisateurs fortement polarisés auraient un apport en diversité plus limité, permettant de les exposer à un contenu progressivement diversifié sans risquer un effet rebond renforçant leur degré de polarisation, tandis que les utilisateurs intermédiaires se verraient apporter une diversité plus importante afin de maintenir leur position intermédiaire et éviter leur enfermement dans une communauté polarisée. Quelle que soit l'approche, **l'adaptation de l'apport en diversité reposant sur une modélisation individuelle et multi-factorielle de polarisation permettrait de personnaliser davantage le service rendu par le système de recommandation, mais aussi de répondre plus directement au phénomène de polarisation.**

6.3.2 Diversification des sentiments exprimés dans les news

Dans le domaine de la recommandation de news, la diversification peut porter sur différents aspects : les sources d'information, les thématiques abordées, *etc.* Dans le cas particulier de la polarisation politique, une diversification des opinions auxquelles sont confrontés les utilisateurs semble primordiale pour promouvoir une prise de décision éclairée. Pour rappel, les approches de diversification exploitent principalement les informations pouvant être extraites des contenus par l'application d'approches de traitement automatique des langues. Pour promouvoir la diversité des opinions, la modélisation des sentiments exprimés dans les news est ainsi nécessaire [Wu *et al.*, 2020a]. Cela permettrait de fournir des recommandations de news exprimant des sentiments variés sur un certain débat, et donc de représenter les différentes opinions existantes sur ce débat.

Des travaux sur cette diversification des sentiments sont en cours de réflexion et résultent, à nouveau, d'une étroite collaboration avec mes collègues du projet ANR BOOM, experts du domaine du traitement automatique des langues. En particulier, ces derniers ont travaillé sur un modèle de classification des sentiments exprimés dans les news [Dufraisse *et al.*, 2022]. Ce modèle a la particularité de fournir des scores de sentiments au niveau des entités, et non pas du document entier. Ces entités sont identifiées suite à l'application d'une approche de reconnaissance d'entités nommées (*Named Entity Recognition*) [Nadeau and Sekine, 2007], et correspondent à des expressions linguistiques pouvant être identifiées de façon unique par un nom propre. Elles correspondent donc à des noms propres décrivant des personnes, des lieux, des organisations, des dates, des valeurs monétaires ou des pourcentages. **La classification des sentiments exprimés sur ces entités, et en particulier celles concernant des personnalités politiques, permettrait donc de caractériser l'opinion exprimée dans un document.**

Plus précisément, le modèle qu'ils ont développé fournit trois scores pour chaque entité identifiée dans une news, liés à trois sentiments : négatif, neutre, positif. Ces scores sur les sentiments pourraient alors être exploités pour caractériser l'opinion d'une news ou d'un ensemble de news. Dans une approche de diversification personnalisée, ces scores permettraient de définir l'opinion d'un utilisateur suivant les news consommées, puis d'adapter les recommandations en conséquence. Cette adaptation des recommandations pourrait notamment reposer sur **la définition d'une nouvelle mesure de similarité entre les items, permettant de quantifier leur proximité sentimentale**. Le calcul du score de pertinence lors de la tâche de recommandation pourrait alors tenir compte d'une proximité thématique, telle que communément appliquée, mais aussi d'une proximité de sentiment. **Cette double évaluation de la proximité des contenus informatifs permettrait ainsi d'affiner le processus de recommandation, en tenant compte d'aspects plus précis des news.** Dans une approche centrée utilisateur, une telle

diversification basée sur les sentiments permettrait une personnalisation d'autant plus fine des recommandations, et garantit donc un contrôle de l'impact de ces recommandations. Cette diversification au niveau des sentiments devrait évidemment répondre aux enjeux éthiques, et ne pas orienter l'opinion des utilisateurs, ni favoriser certaines opinions parmi celles existantes dans le débat public.

6.3.3 Diversifications temporelle et contextuelle

La modélisation temporelle a illustré le caractère évolutif de la polarisation, mais aussi sa forte dépendance au contexte. L'adaptation de la recommandation aux évolutions temporelles et au contexte apparaît alors comme obligatoire.

Dans un premier temps, et dans une perspective de réduction du phénomène de polarisation, la prise en compte des variations temporelles des comportements de polarisation pourrait permettre d'adapter l'apport en diversité. Dans la littérature, ces variations temporelles sont souvent caractérisées comme des dérives conceptuelles (*concept drift*), auxquelles le système doit être en mesure de répondre [Rabiu *et al.*, 2020]. De tels systèmes sont appelés systèmes de recommandation dynamiques, et exploitent des données dynamiques telles que les séries temporelles. L'objectif communément admis de ces systèmes particuliers est de garantir l'exactitude des recommandations au cours du temps, malgré de potentielles variations temporelles des comportements, ou des items. Dans le cadre de la polarisation, c'est également la dimension de diversité qui gagnerait à être adaptée de façon temporelle. Au-delà de la diversité temporelle mise en avant par [Lathia *et al.*, 2010], correspondant à la recommandation d'items différents au cours du temps, c'est l'évolution temporelle du niveau de diversité apporté dans les recommandations que je propose ici. Le niveau de diversité pourrait ainsi être adapté aux périodes de polarisation. Par exemple, lors de l'apparition d'une période équilibrée à la suite d'un événement de contexte, un fort apport en diversité pour les utilisateurs ne prenant pas position (*c.-à-d.* non polarisés) pourrait assurer une prise de décision éclairée. Par ailleurs, lors d'une période de convergence durant laquelle ces utilisateurs intermédiaires tendent à s'approcher des communautés polarisées, le maintien d'un apport en diversité contrôlé pourrait participer à ralentir cette convergence et à maintenir un terrain d'entente entre les communautés opposées. Une adaptation des approches employées par les systèmes de recommandation dynamiques semble alors nécessaire pour répondre à cet objectif.

Dans un second temps, suivant l'évolution de la polarisation à la suite d'événements perturbateurs, une recommandation tenant compte du contexte apparaît intéressante. Les systèmes de recommandation contextuels (*Context-Aware Recommender Systems*) génèrent des recommandations plus exactes en les adaptant au contexte d'utilisation du système par l'utilisateur [Adomavicius and Tuzhilin, 2010]. Ils peuvent par exemple s'adapter au moment de la journée durant lequel l'utilisateur utilise le système (matin, soir, *etc.*), ou encore au type d'appareil utilisé (téléphone portable, tablette, ordinateur, *etc.*). C'est donc le contexte d'utilisation de l'utilisateur qui est pris en compte pour l'adaptation des recommandations. Dans le cas de la polarisation, le contexte auquel le système doit s'adapter est particulier puisqu'il s'agit du contexte politique, ou du débat discuté, durant lequel les recommandations sont fournies à l'utilisateur. Une telle approche constituerait ainsi une nouvelle catégorie spécifique de systèmes de recommandation contextuels. La mise en place d'un tel système nécessite néanmoins d'avoir une connaissance experte du contexte, et donc d'être en mesure d'identifier les événements perturbateurs à l'origine de l'adoption de comportements de polarisation spécifiques. Cette connaissance peut dans certains cas être apportée comme entrée du système, mais nécessite une action volontaire de la part des concepteurs. Dans une perspective d'automatisation, la prédiction précoce d'évé-

nements perturbateurs à partir des interactions des utilisateurs permettrait d'adapter, en temps réel, les recommandations et leur diversification. Pour aller plus loin, l'application de l'apprentissage par transfert (*Transfer Learning*), permettant de transférer les connaissances existantes d'un contexte particulier à un contexte cible [Torrey and Shavlik, 2010], pourrait permettre d'adapter les stratégies lorsque des événements contextuels similaires à d'autres événements déjà observés surgissent.

Pour résumer, **la proposition d'une approche temporelle de recommandation participerait à réduire la polarisation en proposant une diversification contrôlée et progressive des recommandations, tandis que le développement d'approches contextuelles permettrait de s'adapter à la dynamique intrinsèque de cette polarisation.**

6.3.4 Évaluation en ligne du framework ADF

L'évaluation hors ligne du framework ADF a permis de confirmer sa capacité à fournir des recommandations qui sont à la fois exactes, diverses et équitables. Cependant, cette approche d'évaluation hors ligne ne permet pas d'assurer le maintien de ces performances si le framework était appliqué en situation réelle. La mise en place d'une évaluation en ligne du framework ADF permettrait de répondre à ce manque, et de confirmer l'acceptation et l'adoption des recommandations produites par le framework par les utilisateurs réels [Knijnenburg and Willemsen, 2015].

Cette évaluation en ligne est prévue dans le cadre du projet ANR BOOM, au sein duquel j'ai réalisé mes travaux de thèse présentés dans ce manuscrit. La mise en place de cette évaluation en ligne a été discutée avec les différents partenaires du projet. Le protocole d'expérimentation en ligne prévoit ainsi de répartir les participants au sein de différents groupes de taille égale, pour lesquels les recommandations seront fournies par un algorithme de recommandation basé contenu, ou reposant sur une approche de filtrage collaboratif. Parmi ces différents groupes, certains verront leurs recommandations ré-ordonnées suite à l'application du framework ADF. La mise en place de cette évaluation en ligne permettrait ainsi d'évaluer et de comparer les performances de recommandation lorsque le framework ADF est appliqué, ou non. Ce qu'il est intéressant d'évaluer ici c'est notamment la perception qu'ont les utilisateurs vis-à-vis des recommandations qui leurs sont faites. Cette évaluation de la perception peut notamment reposer sur des questionnaires portant sur la satisfaction des utilisateurs, leur perception de la diversité ou de l'équité des recommandations auxquelles ils ont été confrontés, *etc.* Par ailleurs, les résultats d'une telle étude permettraient de déterminer empiriquement les paramètres optimaux du framework ADF pour une bonne acceptabilité des recommandations diversifiées.

Par ailleurs, il a été envisagé de mettre en place cette évaluation en ligne durant un événement politique marquant, lors d'élections par exemple, afin d'évaluer le potentiel impact de la diversification sur la polarisation. Durant ces périodes à fort enjeux, le cadrage médiatique est modifié et est susceptible de modifier la façon dont les utilisateurs accèdent à l'information [Scheufele, 1999]. L'adoption d'une approche personnalisée de diversification est alors essentielle afin de maintenir les utilisateurs informés, sans orienter leurs opinions.

Suivant une approche centrée utilisateur comme celle adoptée tout au long de mon travail de thèse, **la mise en place d'une évaluation en ligne, ou d'une étude utilisateur, permet d'évaluer la satisfaction des utilisateurs finaux du système.** Une optimisation du modèle proposé pour assurer son acceptabilité et sa concordance avec les besoins des utilisateurs peut être proposée en fonction des résulats.

6.3.5 Perspectives à plus long terme

Le développement de systèmes personnalisés de filtrage de l'information est une réponse à l'évolution des pratiques informationnelles suite à la démocratisation d'Internet. Ces systèmes, et notamment les systèmes de recommandation, ont ainsi été largement étudiés dans la littérature et sont désormais utilisés quotidiennement par des millions d'utilisateurs à travers le monde. Néanmoins, il me semble que ces systèmes doivent continuer à évoluer afin de garantir leur pertinence dans un environnement numérique en perpétuelle évolution. Dans le cadre spécifique de la recommandation de news, les pratiques informationnelles ne cessent d'évoluer. Les informations sont désormais transmises sur des médias divers, et de nouveaux vecteurs de l'information apparaissent. Les plateformes telles que Youtube, Tik Tok, Twitch ou Instagram ne sont plus uniquement des plateformes de divertissement, mais deviennent également des sources d'information. Ces dernières sont alors divulguées sous forme de vidéos, le plus souvent très courtes et allant à l'essentiel. Les personnalités politiques elles-mêmes utilisent désormais ces outils pour communiquer sur leurs programmes et idées politiques. L'évolution des pratiques informationnelles s'accompagne donc d'une évolution des supports informationnels, à laquelle les systèmes de recommandation doivent être en mesure de s'adapter. Les approches de recommandation basées contenu, exploitant l'information textuelle des news, sont limitées dans ces nouveaux contextes. Le développement d'approches adaptées à l'exploitation de données de nature différente (vidéos, podcasts, *etc.*), est alors nécessaire [Lubos *et al.*, 2023]. Les approches de diversification imaginées doivent alors elles aussi être adaptées, tout en garantissant toujours une approche personnalisée. **Le développement de systèmes de recommandation doit alors répondre à une évolution constante des pratiques.**

Pour compléter, cette adaptation des systèmes de recommandation doit également tenir compte de l'impact qu'ont ces systèmes sur les utilisateurs. L'IA est plus que jamais au centre des interactions de la société contemporaine, ce qui renforce le besoin de contrôler son impact. Les systèmes développés doivent répondre aux attentes et besoins des utilisateurs, tout en assurant leur liberté et leur sécurité. Leur développement s'accompagne alors d'un respect strict des dimensions éthiques, qu'il me semble essentiel d'évaluer pour en garantir la préservation. Cependant, à l'heure du développement de systèmes profonds dont l'objectif est principalement d'améliorer l'exactitude des recommandations, les cadres d'évaluation communément appliqués peinent à aller au-delà des traditionnelles métriques d'exactitude. En effet, ces systèmes, qui occupent désormais le haut du classement lors de compétitions mises en place sur certains jeux de données, comme MIND [24], ou pour des conférences internationales prestigieuses comme RecSys [25], sont uniquement comparés et évalués au travers de métriques d'exactitude. Les métriques permettant d'évaluer des dimensions allant au-delà de l'exactitude, comme des métriques de diversité, ou encore les métriques évaluant les aspects éthiques ne sont jamais étudiées. Il a pourtant été mis en avant depuis de nombreuses années que l'exactitude seule ne suffit pas à évaluer la qualité d'un système [McNee *et al.*, 2006]. Ces dimensions, qui sont optimisées au même titre que l'exactitude dans le framework ADF, nécessitent selon moi d'être évaluées avec une importance accrue lors de développements de ces nouveaux systèmes de filtrage de l'information. Ces derniers, en plus de prédire au mieux les préférences des utilisateurs, doivent par dessus tout assurer des recommandations répondant aux enjeux sociétaux comme la polarisation en respectant la confidentialité, la transparence, l'autonomie, et l'équité des utilisateurs. Cela nécessite donc la **définition de nouveaux cadres d'évaluation plus stricts et plus complets**, reposant

24. https://msnews.github.io/#leaderboard
25. https://recsys.eb.dk/#leaderboard

sur une évaluation multi-métriques [Jannach and Adomavicius, 2016].

Par ailleurs, le nombre de modèles proposés chaque année, permettant tous d'optimiser les performances de quelques centièmes sur des métriques ad hoc, ne cesse d'augmenter. Je m'interroge sur la pertinence d'une telle course à la performance, et à l'intérêt de développer de si nombreux modèles, répondant pour la plupart à des objectifs communs. Leurs développeurs détiennent une véritable responsabilité quant à leur impact, et notamment dans des contextes sociétaux comme celui de la polarisation. **Une évaluation réglementée et stricte permettrait ainsi de garantir le développement d'une IA responsable.** J'aimerais ajouter que cette responsabilité, en plus de concerner l'impact des systèmes sur les utilisateurs, concerne aussi les aspects écologiques et environnementaux liés au développement de telles approches profondes, nécessitant un apprentissage très gourmand en ressources. D'ici 2040, donc dans un futur proche, les émissions de carbone issues de l'industrie de l'information et de la communication pourraient atteindre jusqu'à 14% des émissions globales au niveau de la planète [Nordgren, 2022]. Un développement raisonné de ces nouvelles approches de recommandation pourrait ainsi participer au développement d'une IA plus verte, essentielle pour répondre à l'urgence climatique.

Finalement, il me paraît également important que les avancées en IA soient accompagnées d'un dialogue avec ses utilisateurs. Dans le cas particulier de la polarisation, abordé dans ce manuscrit, la transparence dans les processus algorithmiques et la participation citoyenne dans la conception des systèmes d'IA sont des éléments intéressants pour garantir que la technologie agisse comme un catalyseur de cohésion sociale plutôt que de division. La mise en place de projets de recherche pluri-disciplinaires, dans le cadre desquels les experts issus de diverses disciplines peuvent confronter leurs idées, et permettant la mise en place d'études à grande échelle impliquant la participation d'utilisateurs réels, participerait ainsi au développement de systèmes mieux adaptés aux enjeux actuels. **La recherche en IA ne se limite plus à des questions techniques, elle est intrinsèquement liée aux défis sociétaux.**

Annexe A

Méthodologie de collecte des données Twitter

Les données étudiées et présentées dans le Chapitre 2 et le Chapitre 3 ont été collectées à l'aide de l'API Twitter (*v2*), avec un accès réservé pour la recherche universitaire. Cette API n'est plus disponible depuis le rachat du réseau social.

La méthodologie de collecte des données repose sur le concept d'utilisateurs élites, qui sont des utilisateurs légitimes pour s'exprimer sur un débat particulier, de par leurs professions, expertises, *etc* [Primario *et al.*, 2017]. Ainsi, pour la collecte des données, il est supposé que les *tweets* postés par ces utilisateurs élites, lorsqu'ils sont liés au débat sélectionné, sont toujours en accord avec leurs convictions.

Les utilisateurs élites sont sélectionnés selon un ensemble de conditions pour garantir leur légitimité :

1. Avoir un nombre significatif de *followers* (*c.-à-d.* le nombre de comptes suivant leur activité) ;
2. Gérer personnellement leur compte Twitter ;
3. Être connus du grand public, grâce à des interventions dans les médias ou au gouvernement ;
4. Être qualifiés, de par leur formation et/ou leur profession, pour traiter un débat ou un sujet particulier.

Les utilisateurs élites constituent un point d'entrée efficace pour la collecte de données sur un sujet spécifique, car leurs opinions sont connues du grand public. Il se distinguent ainsi des autres utilisateurs, appelés utilisateurs standards.

L'objectif est donc d'analyser les comportements d'interaction des utilisateurs standards sur un débat sélectionné et durant une période spécifique. Le jeu de données collecté doit ainsi être équilibré en nombre d'utilisateurs élites appartenant à des communautés opposées, et représentatif des comportements adoptés sur les réseaux sociaux à propos d'un débat spécifique.

Pour obtenir un tel jeu de données, différentes étapes de collecte sont définies après avoir défini le débat étudié, identifié un ensemble pertinent d'utilisateurs élites et défini une période de collecte (Figure A.1). Ces étapes sont les suivantes :

1. Collecte de tous les *tweets* publiés par l'ensemble des utilisateurs élites pendant la période définie ;
2. Filtrage des *tweets* sur le débat étudié ;
3. Collecte d'informations sur un sous-ensemble aléatoire d'utilisateurs standards interagissant pour chaque *tweet* collecté ;

4. Identification des utilisateurs standards les plus actifs parmi ceux sélectionnés à l'étape 3 ;
5. Collecte de toutes les interactions des utilisateurs standards sélectionnés sur les *tweets* des utilisateurs élites collectés au cours de la période définie ;
6. Parmi toutes les interactions collectées, filtrer celles qui sont liées aux *tweets* collectés à l'étape 1.

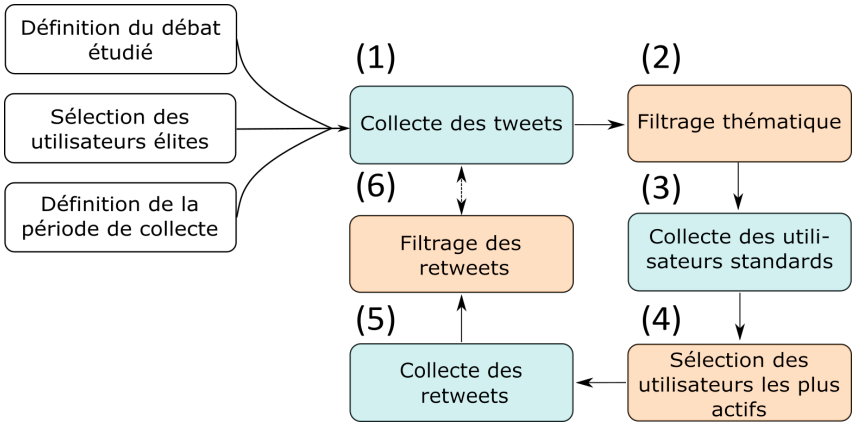


FIGURE A.1 – Étapes de la méthodologie de collecte des données.

Suivant ces étapes, 20 utilisateurs élites francophones ayant une voix légitime dans le débat sur le vaccin contre la COVID-19 (10 pro-vaccins et 10 anti-vaccins), et 20 utilisateurs élites francophones ayant une voix légitime dans le débat sur le conflit en Ukraine (10 pro-Ukraine et 10 pro-Russie) ont été manuellement sélectionnés. Leur opinion est connue car a été clairement exprimée publiquement, et la communauté à laquelle ils se rattachent est donc sans ambiguïté. Afin de préserver leur confidentialité et de respecter la politique de Twitter, les noms ou les noms d'utilisateur des comptes sélectionnés ne sont pas donnés.

Suite à l'identification de ces utilisateurs élites, tous les *tweets* qu'ils ont publiés entre le 1er janvier 2022 et le 31 juillet 2022 ont été collectés. La collecte de données sur une période prolongée (7 mois) permet d'obtenir une quantité suffisante de données de manière à étudier les variations temporelles.

La seconde étape de la méthodologie, consistant à filtrer les *tweets* abordant le débat étudié est basée sur une liste de *hashtags* pertinents liés à chacun des débats étudiés et d'un corpus de *tweets* aléatoires [Turenne, 2018]. Un classifieur à deux classes basé sur le modèle BertTweetFR [Guo *et al.*, 2021] est entraîné sur la base de ces *hashtags* et de ces *tweets*, et permet de ne conserver que les *tweets* des utilisateurs élites traitant ou bien du débat sur le vaccin contre la COVID-19, ou bien du débat sur le conflit en Ukraine.

Enfin, suivant les étapes de méthodologie énoncées, la modélisation de la polarisation détaillée dans ce manuscrit se focalise sur les *retweets*, qui sont des signes d'approbation et donnent donc des informations sur ce avec quoi les utilisateurs sont en accord [Conover *et al.*, 2011]. Ainsi, les informations à propos de 100 *retweeters* choisis au hasard pour chaque *tweet* d'utilisateur élites, ont été collectées. Parmi ces *retweeters* sélectionnés, les 1 000 *retweeters* les plus actifs pour chacun des débats étudiés (500 pro-vaccins et 500 anti-vaccins ; 500 pro-Ukraine et 500 pro-Russie) ont été sélectionnés et sont ceux dont les comportements de polarisation sont modélisés.

Annexe B

Expérimentation préliminaire : évaluation de la diversité des systèmes de recommandation de news

Cette annexe présente un résumé du travail préliminaire mené sur l'évaluation de la diversité dans les recommandations de news, et publié dans le workshop FairUMAP de la conférence UMAP'22.

— **C. Treuillier**, E. Dufraisse, S. Castagnos & A. Brun, (2022) Being Diverse is Not Enough : Rethinking Diversity Evaluation to Meet Challenges of News Recommender Systems, *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP'22)*

B.1 Motivations de recherche

La diversité dans les systèmes de recommandation de news a rapidement été perçue comme un enjeu majeur pour la réduction du phénomène de polarisation. Cet apport en diversité dans les recommandations permettrait en effet de confronter les utilisateurs a un contenu plus divers, participant à ouvrir les bulles d'opinion au sein desquelles ils sont enfermés [Bernstein *et al.*, 2020]. Cependant, l'exposition à des opinions opposées peut, dans certain cas, accroître la polarisation politique [Bail *et al.*, 2018]. Dans ce contexte, il est pertinent de se questionner sur l'impact d'un apport important en diversité dans les recommandations pour tous les utilisateurs. Deux questions de recherche sont ainsi posées :

— **Question *a.*** La diversité des recommandations de news apporte-t-elle un gain systématique ?

— **Question *b.*** L'apport en diversité dans les systèmes de recommandation de news influence-t-il tous les utilisateurs de façon universelle ?

Les travaux menés pour répondre à ces questions de recherche visent à ouvrir le débat sur les pratiques actuelles dans le domaine des médias sociaux et des systèmes de recommandation en ligne.

B.2 Protocole

Pour contribuer à cette réflexion, la diversité des news issues du jeu de données MIND-*small* est étudiée. Ce jeu de données a été collecté durant cinq semaines : les 4 premières pour lesquelles seul l'historique des news consultées par les utilisateurs est fourni, et la $5ème$ pour laquelle la liste des recommandations, détaillant quelles news on été accédées parmi les recommandées, est fournie (voir la Section 5.4.1 pour une description complète).

Pour cette étude préliminaire, un sous-ensemble de données est sélectionné. Seules les interactions effectuées sur les news de la catégorie "*News*" sont sélectionnées, car cette catégorie contient les news politiques. Par ailleurs, seuls les utilisateurs ayant effectué au moins 3 interactions par semaine, sur chacune des 5 semaines de collecte contenu dans le jeu de données MIND-*small* sont sélectionnés. Au final, le jeu de données contient les informations de 1 475 utilisateurs, ayant interagi sur un total de 20 541 news.

La diversité des news est évaluée à l'aide de la métrique de diversité intra-liste (ILD) (Équation 4.10, page 94), qui est la plus communément appliquée dans la littérature. Elle est appliquée sur les *embeddings* des news (voir Annexe C) et permet d'évaluer la diversité de contenu. Les valeurs d'ILD varient entre 0 (pas de diversité) et 1 (diversité maximale). La similarité est calculée à l'aide de la similarité cosinus.

À partir de ces données et de la métrique ILD, quatre diversités différentes sont calculées pour chaque utilisateur :

- La **diversité de l'historique**, correspondant à la diversité moyenne des informations consultées par les utilisateurs au cours des quatre premières semaines de collecte des données.
- La **diversité des recommandations**, correspondant à la diversité moyenne des news recommandées au cours de la $5^{ème}$ semaine.
- La **diversité des news accédées**, correspondant à la diversité moyenne des news consultées par les utilisateurs au cours de la $5^{ème}$ semaine parmi les recommandations.
- La **diversité des news non accédées**, correspondant à la diversité moyenne des news recommandées à chaque utilisateur, mais non consultées.

Ces diversités sont dans un premier temps étudiées dans le cadre d'une (1) **analyse holistique**. Cette dernière permet d'étudier la répartition des valeurs de diversité globalement, sur les cinq semaines de collecte de données. Dans un second temps et pour compléter cette analyse holistique, une (2) **analyse temporelle** est proposée. Les utilisateurs sont divisés au sein de quatre groupes d'utilisateurs de taille égale. Ces groupes sont définis en fonction des quartiles identifiés à partir des valeurs de diversité moyenne des news accédées par chaque utilisateur chaque semaine. Le premier quart contient les utilisateurs dont les valeurs de diversité sont les plus faibles. Les variations temporelles sont alors analysées en fonction des variations de quartiles entre les utilisateurs, c'est à dire le passage d'un quartile à un autre.

B.3 Résultats

B.3.1 Analyse holistique

La Figure B.1 présente la distribution des différentes valeurs de diversité pour l'ensemble des utilisateurs. L'analyse de ces distributions permet tout d'abord d'observer que la diversité

moyenne des recommandations (courbe rouge Figure B.1) est élevée $(0,75)$, et que l'écart type est réduit $(0,04)$. Cette diversité moyenne élevée et le faible écart-type des news recommandées indique qu'un processus de diversification est mis en place dans le système de recommandation de la plateforme d'actualités de Microsoft, d'où sont issues les données. En particulier, cet algorithme semble garantir un niveau de diversité minimal dans les recommandations fournies aux utilisateurs. Je précise également que la distribution des news non accédées (courbe verte dans la Figure B.1) est très proche de celle des recommandations car une large partie des recommandations n'est pas accédée (93%).

L'analyse de la distribution des valeurs de diversité pour les news consommées (courbe orange Figure B.1) permet tout d'abord d'observer une différence importante avec celles des recommandations. La diversité moyenne des news consommées est notamment nettement inférieure à celle des news recommandées. Dans les faits, $67,1\%$ des utilisateurs accèdent à des news moins diverses que leurs recommandations. Cette distribution des valeurs de diversité des news accédées s'accompagne d'un écart-type plus élevé que pour les recommandations $(0,17$ *vs.* $0,04)$, reflétant une plus grande variabilité de la diversité moyenne des news consultées par les utilisateurs. Finalement, la distribution des valeurs de diversité dans les news de l'historique est similaire à celle des news accédées.

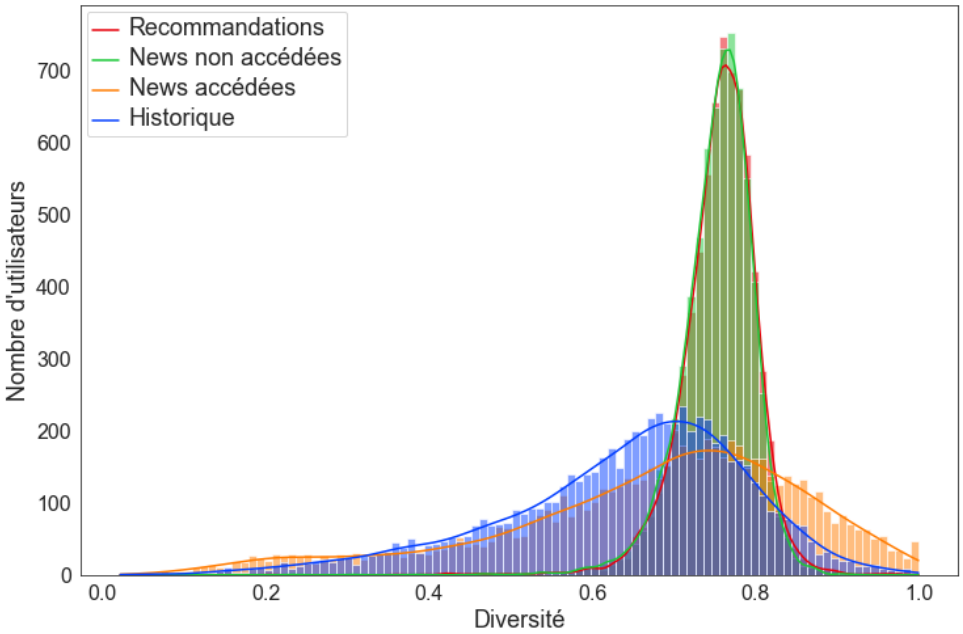


FIGURE B.1 – Distributions des diversités.

Cette analyse holistique fournit des premiers éléments permettant d'affirmer qu'**une importante diversité de recommandations ne conduit pas systématiquement à une consommation de news diversifiées**.

B.3.2 Analyse temporelle

La figure B.2 représente le flux entre les quartiles d'une semaine à l'autre dans un diagramme de Sankey. L'épaisseur de chaque flux est proportionnelle au nombre d'utilisateurs dans ce flux. L'analyse de ce diagramme permet de souligner que les schémas de transition restent stables au fil des semaines. Il y a donc un impact global similaire des recommandations sur la diversité des

news consultées par les utilisateurs au fil des semaines. Par ailleurs, chaque transition possible (de n'importe quel quartile à n'importe quel autre quartile) se produit chaque semaine. Cependant, près de 80% des utilisateurs restent dans le même quartile ou dans un quartile adjacent entre deux semaines consécutives. Ainsi, peu d'utilisateurs changent significativement leurs habitudes de consommation entre deux semaines.

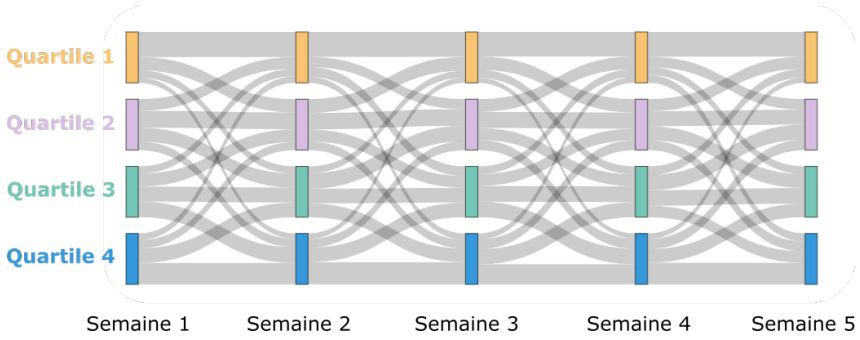


FIGURE B.2 – Diagramme de Sankey montrant les évolutions de quartiles de semaine en semaine.

Cette étude des variations temporelle reste limitée, et ne permet pas d'analyser finement les potentiels comportements temporels spécifiques adoptés par les utilisateurs. Pour affiner cette analyse, et identifier de potentielles variations temporelles spécifiques, je propose d'évaluer les variations de quartiles chaque semaine par rapport à la première semaine pour chaque utilisateur. Ces variations sont quantifiées par un entier appartenant à l'intervalle $[-3, +3]$. Une variation est égale à 3 lorsqu'un utilisateur passe du 1^{er} au $4^{ème}$ quartile, *c.-à-d.* que sa diversité augmente fortement. Une variation égale à -3 représente un utilisateur qui passe du $4^{ème}$ au 1^{er}, *c.-à-d.* que sa diversité diminue considérablement. Si la variation est égale à 0, l'utilisateur reste dans le même quartile.

Les variations de quartiles sont représentées visuellement par une carte thermique (*heatmap*) dans la Figure B.3. Une variation rouge représente une augmentation de la diversité, plus elle est foncée, plus elle est importante. Une variation en bleu représente une diminution de la diversité, plus elle est foncée, plus elle est importante.

Cette figure fournit une vision globale des variations de chaque utilisateur par rapport à leur diversité moyenne au cours de la première semaine. L'analyse de la colonne la plus à droite permet d'identifier trois types d'utilisateurs en fonction de leur réceptivité à la diversité des recommandations.

— **Utilisateurs positivement réceptifs** : utilisateurs accédant à des informations plus diversifiées après 5 semaines d'utilisation du système.

— **Utilisateurs négativement réceptifs** : utilisateurs accédant à des informations moins diversifiées après 5 semaines d'utilisation du système.

— **Utilisateurs résistants** : utilisateurs accédant à des informations tout aussi diverses après 5 semaines d'utilisation du système.

Plus précisément, ces trois types d'utilisateurs sont divisés en trois tiers : $32,1\%$ des utilisateurs sont positivement réceptifs (nuances de rouge dans la dernière colonne de la Figure B.3), $32,5\%$ des utilisateurs sont négativement réceptifs (nuances de bleu dans la dernière colonne de la Figure B.3), et $35,4\%$ des utilisateurs sont résistants à la diversité (bande blanche dans la dernière colonne de la Figure B.3). L'apport en diversité élevé pour l'ensemble des utilisateurs semble donc pouvoir avoir un impact négatif, neutre ou positif avec la même probabilité.

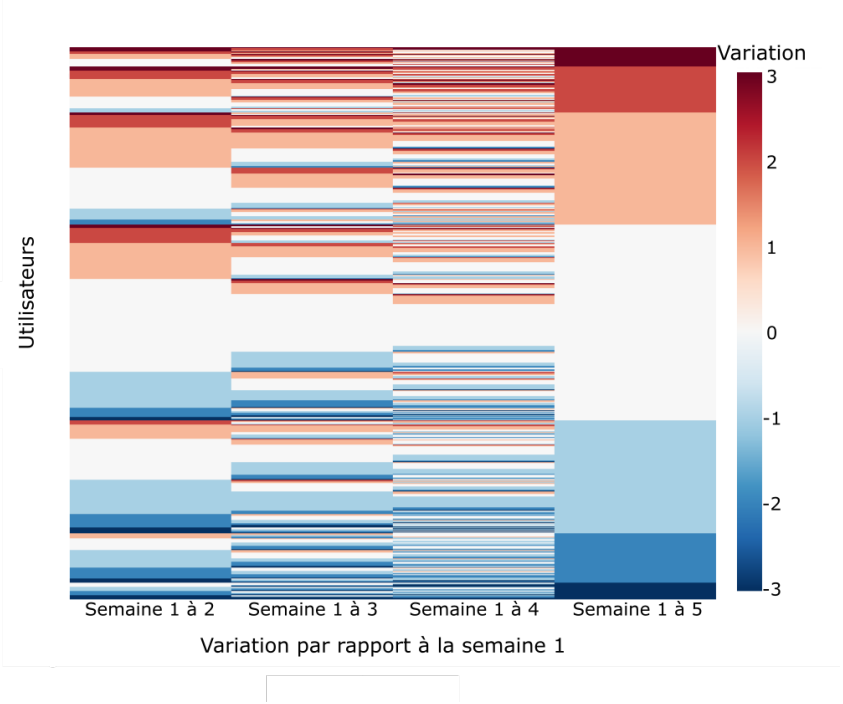


FIGURE B.3 – Carte thermique des changements de quartiles entre la semaine 1 et toutes les autres semaines. Chaque ligne pixelisée représente un utilisateur.

B.4 Conclusion

Cette analyse préliminaire de la diversité dans les recommandations de news a permis dans un premier temps de mettre en avant qu'un apport élevé et commun à tous les utilisateurs en diversité ne garantit pas une consommation de news diversifiée de façon systématique (**Question *a***). Dans un second temps, l'analyse temporelle a permis de mettre en avant que cette diversité élevée des recommandations a un impact différent en fonction des utilisateurs (**Question *b***). Plus précisément, pour autant d'utilisateurs, cet apport en diversité peut être ou bien positif et entraîner une augmentation de la diversité dans les consommations, ou bien négatif et entraîner une diminution de la diversité dans les consommations, ou bien neutre et ne pas influencer la diversité des consommations. Ces conclusions tendent ainsi à confirmer celles de [Bail *et al.*, 2018], selon lesquelles exposer des utilisateurs polarisés à la diversité peut s'avérer contre-productif. La diversité n'est pas accueillie de façon universelle par l'ensemble des utilisateurs. Ainsi, **la distinction de différents profils d'utilisateurs distingués par l'impact de l'apport en diversité sur leur consommation permet de confirmer le besoin d'approches de recommandation, et en particulier de diversification, qui soient personnalisées et adaptées aux besoins de chaque utilisateur.**

Annexe C

Méthodologie de représentation numérique des news

Dans un premier temps, l'ensemble du contenu des news disponibles dans le jeu de données MIND-*large* a été extrait à partir des URLs fournies, à l'aide de la librairie Trafilatura [26].

Dans un second temps, les contenus extraits ont été numériquement représentés suivant une approche non supervisée : le modèle LDA (Latent Dirichlet Allocation) [Blei *et al.*, 2003], permettant de modéliser la distribution thématique des news, a été appliqué. L'idée de ce modèle est de décomposer un ensemble de contenus en thèmes sous-jacents. La librairie utilisée pour l'application du modèle LDA sur le contenu des news disponibles est la librairie Tomotopy [27].

Les différentes étapes préliminaires à l'application du modèle LDA, une fois leur contenu extrait, sont les suivantes :

- Concaténation de l'ensemble des contenus originaux des ensembles d'apprentissage, de validation et de test du jeu de données MIND.
- Dé-duplication des contenus disponibles à l'aide de la méthode MinMashLSH, reposant sur la similarité de Jaccard pour identifier les documents dupliqués. En sortie, 126 649 news distinctes sont fournies.
- Pré-traitement des contenus avec retrait des *stop-words*, retrait des chiffres, et lemmatisation des mots.
- Formation de représentation des contenus sous forme de sacs de mots (*bag-of-words*).
- Filtrage des news contenant moins de 20 mots. Les news conservées contiennent ainsi entre 21 et 4 351 mots, avec une moyenne de 261 mots.

Une fois les contenus des news pré-traités, le modèle LDA est appliqué. Les hyperparamètres α, contrôlant la distribution de Dirichlet pour les documents, η, contrôlant la distribution de Dirichlet pour les mots, et k, représentant le nombre de thématiques permettant de représenter les contenus, ont été optimisés. Plus précisément, les valeurs de $\alpha \in [1e{-}4; 0, 1]$, $\eta \in [1e^{-4}; 0, 1]$ et $k \in \{32, 64, 128\}$ ont été testées. Cette optimisation repose sur la mesure de perplexité.

L'application de ce modèle LDA, avec des hyper-paramètres optimisés, a ainsi permis de fournir une représentation numérique de l'ensemble des news du jeu de données MIND-large sous forme d'*embeddings* de 128 dimensions.

26. https://trafilatura.readthedocs.io/en/latest/
27. https://bab2min.github.io/tomotopy/v0.12.2/en/

Annexe D

Méthodologie de modélisation thématique des news

La modélisation thématique (*c.-à-d.* l'identification des différents sujets abordés dans les news) repose sur l'exploitation des *embeddings* LDA des news.

Cette modélisation a été appliquée suivant une méthodologie commune du domaine du traitement automatique des langues, consistant à appliquer l'algorithme UMAP [McInnes *et al.*, 2018] afin de réduire la dimensionnalité des *embeddings*, puis à appliquer l'algorithme de clustering HDBSCAN [Campello *et al.*, 2013] afin d'identifier les clusters correspondant à des thématiques distinctes.

Les paramètres optimisables de l'algorithme UMAP, appliqués à l'aide de la librairie umap-learn [28] sont les suivants :

- *n_neighbors* : taille du voisinage utilisé pour l'approximation. Les valeurs sont généralement comprises entre 2 et 100.
- *min_dist* : distance minimale effective entre deux points des *embeddings*.
- *n_components* : nombre de dimensions attendues dans la représentation réduite résultante.

Il est important de noter que le paramètre *min_dist* est fixé à 0 lorsque l'algorithme UMAP est utilisé avant l'application d'un algorithme de clustering comme HDBSCAN. Ce paramètre n'a donc pas été optimisé dans le cadre de la modélisation thématique des news du jeu de données MIND.

Les paramètres optimisables de l'algorithme HDBSCAN, appliqué à l'aide de la librairie hdbscan [29], sont les suivants :

- *min_cluster_size* : taille minimale d'un cluster.
- *min_samples* : nombre d'échantillons dans le voisinage d'un point pour que celui-ci soit considéré comme un point central.

Les valeurs de ces différents paramètres qui ont été testées pour l'optimisation sont présentées dans le Tableau D.1.

Les valeurs de ces différents paramètres ont été optimisées de façon à maximiser deux indices de performances : l'indice de Silhouette [Rousseeuw, 1987] et l'indice DBCV [Moulavi *et al.*, 2014]. Les valeurs de ces deux indices de performance varient dans $[-1; 1]$, avec des valeurs élevées indiquant des performances accrues.

28. https://umap-learn.readthedocs.io/en/latest/index.html
29. https://hdbscan.readthedocs.io/en/latest/index.html

Paramètre	Valeurs
n_ neighbors	30, 100
min_ dist	0
n_ components	2, 4, 6, 8, 10
min_ cluster_ size	100, 200, 300
min_ samples	1, 10, 50

TABLE D.1 – Valeurs des paramètres testées pour l'optimisation.

Les performances optimales ont été obtenues avec *n_ neighbors*= 30, *n_ components*= 4, *min_ cluster_ size*= 300 et *min_ samples*= 50, permettant d'avoir un indice de Silhouette de $0,23$ et un score DBCV de $0,60$. L'application des algorithmes UMAP et HDBSCAN avec ces paramètres permet d'identifier 13 clusters, correspondant à 13 sujets (*topics*) représentant les aspects caractérisant les news du jeu de données MIND sur lesquelles le framework ADF est appliqué.

Annexe E

Résultats détaillés de la validation expérimentale du framework ADF

Cette annexe présente les résultats chiffrés des expérimentations présentées dans la Section 5.4 de ce manuscrit. Dans chacun des tableaux, les valeurs en gras correspondent aux performances maximales pour la métrique considérée ($Prcision$, ILD, $S-Recall$ ou C_H).

$Précision$@20	ILD	$S-Recall$@20	C_H
0,224	0,450	0,383	0,364

Table E.1 – Performances avec la A-Baseline

$Précision$@20	ILD	$S-Recall$@20	C_H
0,227	0,562	0,525	0,141

Table E.2 – Performances avec la AF-Baseline

Valeur de λ	*Précision*@20	ILD	$S-Recall$@20	C_H
$\lambda = 0$	0,224	0,450	0,383	0,364
$\lambda = 0,1$	0,224	0,450	0,383	0,360
$\lambda = 0,2$	0,224	0,451	0,383	0,359
$\lambda = 0,3$	**0,225**	0,452	0,384	0,359
$\lambda = 0,4$	**0,225**	0,455	0,386	0,356
$\lambda = 0,5$	0,224	0,472	0,395	0,349
$\lambda = 0,6$	0,223	0,481	0,399	0,348
$\lambda = 0,7$	0,223	0,492	0,406	0,344
$\lambda = 0,8$	0,222	0,525	0,427	0,335
$\lambda = 0,9$	0,217	0,614	0,492	**0,316**
$\lambda = 1$	0,173	**0,848**	**0,613**	0,370

TABLE E.3 – Performances avec la AD-Baseline

Valeur de α	*Précision*@20	ILD	$S-Recall$@20	C_H
$\alpha = 0$	**0,227**	0,562	0,525	0,141
$\alpha = 0,1$	0,225	0,570	0,552	0,038
$\alpha = 0,2$	0,225	0,572	0,559	0,030
$\alpha = 0,3$	0,222	0,581	0,588	0,016
$\alpha = 0,4$	0,220	0,593	0,622	0,055
$\alpha = 0,5$	0,217	0,602	0,646	0,089
$\alpha = 0,6$	0,215	0,607	0,661	0,110
$\alpha = 0,7$	0,176	0,702	**1,000**	0,072
$\alpha = 0,8$	0,171	0,709	**1,000**	0,041
$\alpha = 0,9$	0,162	0,724	**1,000**	0,003
$\alpha = 1$	0,123	**0,794**	**1,000**	**0,011**

TABLE E.4 – Performances avec ADF

Diversité cible globale	*Précision*@20	ILD	$S-Recall$@20	C_H
diversité$= 0,6$	**0,226**	0,567	0,528	0,114
diversité$= 0,7$	0,223	0,577	0,558	0,047
diversité$= 0,8$	0,211	0,615	0,660	0,138
diversité$= 0,9$	0,175	0,702	0,930	0,042
diversité$= 1$	0,123	**0,794**	**0,945**	**0,011**

TABLE E.5 – Performances lors d'un apport en diversité global avec ADF

Groupe	Nombre d'utilisateurs	*Diversité de sélection*
Groupe 1	3	$[0,1;0,2[$
Groupe 2	24	$[0,2;0,3[$
Groupe 3	133	$[0,3;0,4[$
Groupe 4	525	$[0,4;0,5[$
Groupe 5	1964	$[0,5;0,6[$
Groupe 6	4250	$[0,6;0,7[$
Groupe 7	2894	$[0,7;0,8[$
Groupe 8	207	$[0,8;0,9]$

TABLE E.6 – Détails sur les groupes d'utilisateurs formés

Groupe	*Précision*@20	ILD	$S-Recall$@20
Groupe 1	0,122	**0,742**	**0,974**
Groupe 2	0,142	0,642	0,901
Groupe 3	0,190	0,667	0,891
Groupe 4	0,190	0,667	0,872
Groupe 5	**0,218**	0,584	0,637
Groupe 6	0,215	0,615	0,630
Groupe 7	0,210	0,636	0,662
Groupe 8	0,172	0,661	0,711

TABLE E.7 – Performances pour chacun des groupes lorsque l'apport global en diversité est de $0,8$

Groupe	*Précision*@20	ILD	$S-Recall$@20
Groupe 1	0,239	0,422	0,256
Groupe 2	0,208	0,324	0,394
Groupe 3	**0,278**	0,417	0,404
Groupe 4	0,258	0,465	0,456
Groupe 5	0,236	0,533	0,526
Groupe 6	0,217	0,600	0,613
Groupe 7	0,207	0,651	0,712
Groupe 8	0,165	**0,689**	**0,813**

TABLE E.8 – Performances pour chacun des groupes lorsque l'apport en diversité est personnalisé, avec $\alpha = 0,6$

Bibliographie

[Abdollahpouri *et al.*, 2020] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. The connection between popularity bias, calibration, and fairness in recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 726–731, 2020.

[Abramowitz and Saunders, 2008] Alan I Abramowitz and Kyle L Saunders. Is polarization a myth ? *The Journal of Politics*, 70(2) :542–555, 2008.

[Abrams, 2007] Samuel J Abrams. Polarized america : The dance of ideology and unequal riches. *Perspectives on Politics*, 5(3) :642–644, 2007.

[Adams *et al.*, 2022] Johnathan A Adams, Gentry White, and Robyn P Araujo. Mathematical measures of societal polarisation. *Plos one*, 17(10) :e0275283, 2022.

[Adomavicius and Kwon, 2011] Gediminas Adomavicius and YoungOk Kwon. Maximizing aggregate recommendation diversity : A graph-theoretic approach. In *Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), held in conjunction with ACM RecSys'11*, pages 3–10, Chicago, USA, 2011. ACM.

[Adomavicius and Tuzhilin, 2005] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6) :734–749, 2005.

[Adomavicius and Tuzhilin, 2010] Gediminas Adomavicius and Alexander Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2010.

[Aggarwal, 2016] Charu C. Aggarwal. *Recommender Systems - The Textbook*. Springer, Berlin, Germany, 2016.

[Agrawal *et al.*, 2009] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the second ACM international conference on web search and data mining*, pages 5–14, 2009.

[Alabduljabbar *et al.*, 2023] Reham Alabduljabbar, Halah Almazrou, and Amaal Aldawod. Context-aware news recommendation system : Incorporating contextual information and collaborative filtering techniques. *International Journal of Computational Intelligence Systems*, 16(1) :137, 2023.

[Alvim *et al.*, 2023] Mário S Alvim, Bernardo Amorim, Sophia Knight, Santiago Quintero, and Frank Valencia. A formal model for polarization under confirmation bias in social networks. *Logical Methods in Computer Science*, 19, 2023.

[An *et al.*, 2019] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 336–345, 2019.

[Arrègle, 2003] Jean-Luc Arrègle. Les modèles linéaires hiérarchiques : 1. principes et illustration. *M@ n@ gement*, 6(1) :1–28, 2003.

[Azzimonti and Fernandes, 2023] Marina Azzimonti and Marcos Fernandes. Social media networks, fake news, and polarization. *European journal of political economy*, 76 :102256, 2023.

[Badami *et al.*, 2017] Mahsa Badami, Olfa Nasraoui, Welong Sun, and Patrick Shafto. Detecting polarization in ratings : An automated pipeline and a preliminary quantification on several benchmark data sets. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 2682–2690. IEEE, 2017.

[Baeza-Yates *et al.*, 1999] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[Bail *et al.*, 2018] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37) :9216–9221, 2018.

[Balabanović and Shoham, 1997] Marko Balabanović and Yoav Shoham. Fab : content-based, collaborative recommendation. *Communications of the ACM*, 40(3) :66–72, 1997.

[Baldassarri and Bearman, 2007] Delia Baldassarri and Peter Bearman. Dynamics of political polarization. *American sociological review*, 72(5) :784–811, 2007.

[Baracskay *et al.*, 2022] Ian Baracskay, Donald J Baracskay III, Mehtab Iqbal, and Bart Piet Knijnenburg. The diversity of music recommender systems. In *Proc. of the International conference IUI 2022*, IUI '22 Companion, page 97–100, Helsinki, Finland, 2022. ACM.

[Barberá, 2020] Pablo Barberá. Social media, echo chambers, and political polarization. *Social media and democracy : The state of the field, prospects for reform*, 34, 2020.

[Barkan *et al.*, 2023] Oren Barkan, Tom Shaked, Yonatan Fuchs, and Noam Koenigstein. Modeling users' heterogeneous taste with diversified attentive user profiles. *User Modeling and User-Adapted Interaction*, pages 1–31, 2023.

[Batmaz *et al.*, 2019] Zeynep Batmaz, Ali Yurekli, Alper Bilge, and Cihan Kaleli. A review on deep learning for recommender systems : challenges and remedies. *Artificial Intelligence Review*, 52 :1–37, 2019.

[Baumann *et al.*, 2020] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4) :048301, 2020.

[Baumann *et al.*, 2021] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. Emergence of polarized ideological opinions in multidimensional topic spaces. *Physical Review X*, 11(1) :011012, 2021.

[Baxter, 2016] Rodney J Baxter. *Exactly solved models in statistical mechanics*. Elsevier, 2016.

[Beam *et al.*, 2020] Michael A Beam, Myiah J Hutchens, and Jay D Hmielowski. Facebook news and (de) polarization : Reinforcing spirals in the 2016 us election. In *Digital media, political polarization and challenges to democracy*, pages 26–44. Routledge, 2020.

[Beauguitte, 2010] Laurent Beauguitte. Graphes, réseaux, réseaux sociaux : vocabulaire et notation. 2010.

[Becatti *et al.*, 2019] Carolina Becatti, Guido Caldarelli, Renaud Lambiotte, and Fabio Saracco. Extracting significant signal of news consumption from social networks : the case of twitter in italian political elections. *Palgrave Communications*, 5(1) :1–16, 2019.

[Beel and Langer, 2015] Joeran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In *Research and Advanced Technology for Digital Libraries : 19th International Conference on Theory and Practice of Digital Libraries, TPDL 2015, Poznań, Poland, September 14-18, 2015, Proceedings 19*, pages 153–168. Springer, 2015.

[Belkin and Croft, 1992] Nicholas J Belkin and W Bruce Croft. Information filtering and information retrieval : Two sides of the same coin ? *Communications of the ACM*, 35(12) :29–38, 1992.

[Bellogin *et al.*, 2011] Alejandro Bellogin, Pablo Castells, and Ivan Cantador. Precision-oriented evaluation of recommender systems : an algorithmic comparison. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 333–336, 2011.

[Bernstein *et al.*, 2020] Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A Beam, Marc P Hauer, Lucien Heitz, Pascal Jürgens, et al. Diversity in news recommendations. *arXiv preprint arXiv :2005.09495*, 2020.

[Bessi *et al.*, 2016] Alessandro Bessi, Fabiana Zollo, Michela Del Vicario, Michelangelo Puliga, Antonio Scala, Guido Caldarelli, Brian Uzzi, and Walter Quattrociocchi. Users polarization on facebook and youtube. *PloS one*, 11(8) :e0159641, 2016.

[Biega *et al.*, 2018] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. Equity of attention : Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 405–414, New York, NY, USA, 2018. Association for Computing Machinery.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3(4-5) :993–1022, 2003.

[Bonicalzi *et al.*, 2023] Sofia Bonicalzi, Mario De Caro, and Benedetta Giovanola. Artificial intelligence and autonomy : on the ethical dimension of recommender systems. *Topoi*, 42(3) :819–832, 2023.

[Bourque *et al.*, 2009] Jimmy Bourque, Jean-Guy Blais, and François Larose. L'interprétation des tests d'hypothèses : p, la taille de l'effet et la puissance. *Revue des sciences de l'éducation*, 35(1) :211–226, 2009.

[Boyadjian *et al.*, 2014] Julien Boyadjian, Marie Neihouser, MM in Skoric, P Parycek, and M Sachs. Why and how to create a panel of twitter users. In *CeDEM Asia où les uti2014 : Conference for E-Democracy an Open Government*, pages 247–252. Donau-Universitãt Krems Krems, 2014.

[Bozdag and van den Hoven, 2015] Engin Bozdag and Jeroen van den Hoven. Breaking the filter bubble : democracy and design. *Ethics and Information Technology*, 17(4) :249–265, 2015.

[Bozdag, 2013] Engin Bozdag. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15 :209–227, 2013.

[Bradley and Smyth, 2001] Keith Bradley and Barry Smyth. Improving recommendation diversity. In *Proceedings of the twelfth Irish conference on artificial intelligence and cognitive science*, volume 85, pages 141–152, Maynooth, Ireland, 2001. Citeseer, NUIM Department of Computer Science.

[Bruns, 2019] Axel Bruns. It's not the technology, stupid : How the 'echo chamber'and 'filter bubble'metaphors have failed us. *International Association for Media and Communication Research*, 2019.

[Buder *et al.*, 2021] Jürgen Buder, Lisa Rabl, Markus Feiks, Mandy Badermann, and Guido Zurstiege. Does negatively toned language use on social media lead to attitude polarization ? *Computers in Human Behavior*, 116 :106663, 2021.

[Burke, 2002] Robin Burke. Hybrid recommender systems : Survey and experiments. *User modeling and user-adapted interaction*, 12 :331–370, 2002.

[Cagé *et al.*, 2020] Julia Cagé, Nicolas Hervé, and Marie-Luce Viaud. The production of information in an online world. *The Review of Economic Studies*, 87(5) :2126–2164, 2020.

[Campello *et al.*, 2013] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[Çano and Morisio, 2017] Erion Çano and Maurizio Morisio. Hybrid recommender systems : A systematic literature review. *Intelligent data analysis*, 21(6) :1487–1524, 2017.

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

[Carothers and O'Donohue, 2019] Thomas Carothers and Andrew O'Donohue. *Democracies divided : The global challenge of political polarization*. Brookings Institution Press, 2019.

[Carpini and Keeter, 1996] Michael X Delli Carpini and Scott Keeter. *What Americans know about politics and why it matters*. Yale University Press, 1996.

[Chang and Park, 2021] Kiyoung Chang and Jeeyoung Park. Social media use and participation in dueling protests : The case of the 2016–2017 presidential corruption scandal in south korea. *The International Journal of Press/Politics*, 26(3) :547–567, 2021.

[Chen *et al.*, 2023] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. Bias and debias in recommender system : A survey and future directions. *ACM Transactions on Information Systems*, 41(3) :1–39, 2023.

[Chicco *et al.*, 2021] Davide Chicco, Matthijs J. Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smap, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7, 2021.

[Cicchini *et al.*, 2022] Tomas Cicchini, Sofia Morena Del Pozo, Enzo Tagliazucchi, and Pablo Balenzuela. News sharing on twitter reveals emergent fragmentation of media agenda and persistent polarization. *EPJ Data Science*, 11(1) :48, 2022.

[Clarke *et al.*, 2008] Charles LA Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666, 2008.

[Clauset *et al.*, 2004] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6) :066111, 2004.

[Claypool *et al.*, 1999] Mark Claypool, Anuja Gokhale, Tim Miranda, Paul Murnikov, Dmitry Netes, and Matthew Sartin. Combing content-based and collaborative filters in an online newspaper. In *Proc. of Workshop on Recommender Systems-Implementation and Evaluation*, 1999.

[Conover *et al.*, 2011] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In

Proceedings of the international aaai conference on web and social media, volume 5, pages 89–96, 2011.

[Conroy-Krutz and Moehler, 2015] Jeffrey Conroy-Krutz and Devra C Moehler. Moderation from bias : A field experiment on partisan media in a new democracy. *The Journal of Politics*, 77(2) :575–587, 2015.

[da Silva *et al.*, 2021] Diego Corrêa da Silva, Marcelo Garcia Manzato, and Frederico Araújo Durão. Exploiting personalized calibration and metrics for fairness recommendation. *Expert Systems with Applications*, 181 :115112, 2021.

[Dandekar *et al.*, 2013] Pranav Dandekar, Ashish Goel, and David T Lee. Biased assimilation, homophily, and the dynamics of polarization. *Proceedings of the National Academy of Sciences*, 110(15) :5791–5796, 2013.

[Davies and Bouldin, 1979] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2) :224–227, 1979.

[de Campos *et al.*, 2023] Luis M de Campos, Juan M Fernández-Luna, and Juan F Huete. Use of topical and temporal profiles and their hybridisation for content-based recommendation. *User Modeling and User-Adapted Interaction*, pages 1–27, 2023.

[de Souza Pereira Moreira *et al.*, 2018] Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. News session-based recommendations using deep neural networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, pages 15–23, Vancouver, Canada, 2018. ACM.

[DeGroot, 1974] Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345) :118–121, 1974.

[Del Vicario *et al.*, 2016] Michela Del Vicario, Gianna Vivaldo, Alessandro Bessi, Fabiana Zollo, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Echo chambers : Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1) :37825, 2016.

[Di Noia *et al.*, 2014] Tommaso Di Noia, Vito Claudio Ostuni, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. An analysis of users' propensity toward diversity in recommendations. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 285–288, 2014.

[Diakopoulos and Koliska, 2017] Nicholas Diakopoulos and Michael Koliska. Algorithmic transparency in the news media. *Digital journalism*, 5(7) :809–828, 2017.

[Druckman *et al.*, 2013] James N Druckman, Erik Peterson, and Rune Slothuus. How elite partisan polarization affects public opinion formation. *American political science review*, 107(1) :57–79, 2013.

[Dufraisse *et al.*, 2022] Evan Dufraisse, Célina Treuillier, Armelle Brun, Julien Tourille, Sylvain Castagnos, and Adrian Popescu. Don't burst blindly : For a better use of natural language processing to fight opinion bubbles in news recommendations. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 79–85, 2022.

[Eger, 2016] Steffen Eger. Opinion dynamics and wisdom under out-group discrimination. *Mathematical Social Sciences*, 80 :97–107, 2016.

[Ekstrand *et al.*, 2011] Michael D Ekstrand, John T Riedl, Joseph A Konstan, et al. Collaborative filtering recommender systems. *Foundations and Trends® in Human–Computer Interaction*, 4(2) :81–173, 2011.

[Elahi *et al.*, 2022] Mehdi Elahi, Dietmar Jannach, Lars Skjærven, Erik Knudsen, Helle Sjøvaag, Kristian Tolonen, Øyvind Holmstad, Igor Pipkin, Eivind Throndsen, Agnes Stenbom, et al. Towards responsible media recommendation. *AI and Ethics*, pages 1–12, 2022.

[Eliasoph, 1998] Nina Eliasoph. *Avoiding politics : How Americans produce apathy in everyday life*. Cambridge University Press, 1998.

[Eskandanian *et al.*, 2017] Farzad Eskandanian, Bamshad Mobasher, and Robin Burke. A clustering approach for personalizing diversity in collaborative recommender systems. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 280–284, 2017.

[Esteban and Ray, 1994] Joan-Maria Esteban and Debraj Ray. On the measurement of polarization. *Econometrica : Journal of the Econometric Society*, pages 819–851, 1994.

[Felfernig *et al.*, 2015] Alexander Felfernig, Gerhard Friedrich, Dietmar Jannach, and Markus Zanker. Constraint-based recommender systems. *Recommender systems handbook*, pages 161–190, 2015.

[Ferdousi *et al.*, 2017] Zahra Vahidi Ferdousi, Elsa Negre, and Dario Colazzo. Context factors in context-aware recommender systems. In *AISR 2017 : Atelier interdisciplinaire sur les systèmes de recommandation*, 2017.

[Ferrari Dacrema *et al.*, 2019] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress ? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM conference on recommender systems*, pages 101–109, 2019.

[Festinger, 1962] Leon Festinger. Cognitive dissonance. *Scientific American*, 207(4) :93–106, 1962.

[Fiorina and Abrams, 2008] Morris P Fiorina and Samuel J Abrams. Political polarization in the american public. *Annu. Rev. Polit. Sci.*, 11 :563–588, 2008.

[Fkih, 2022] Fethi Fkih. Similarity measures for collaborative filtering-based recommender systems : Review and experimental comparison. *Journal of King Saud University-Computer and Information Sciences*, 34(9) :7645–7669, 2022.

[Fletcher and Nielsen, 2018] Richard Fletcher and Rasmus Kleis Nielsen. Are people incidentally exposed to news on social media ? a comparative analysis. *New media & society*, 20(7) :2450–2468, 2018.

[Floridi, 2011] Luciano Floridi. The construction of personal identities online. *Minds and Machines*, 21(4) :477–479, 2011.

[Garcia *et al.*, 2015] David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer. Ideological and temporal components of network polarization in online political participatory media. *Policy & internet*, 7(1) :46–79, 2015.

[Garimella and Weber, 2017] Venkata Rama Kiran Garimella and Ingmar Weber. A long-term analysis of polarization on twitter. In *Proceedings of the International AAAI Conference on Web and social media*, volume 11, pages 528–531, 2017.

[Garimella *et al.*, 2018] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1) :1–27, 2018.

[Garimella *et al.*, 2021] Kiran Garimella, Tim Smith, Rebecca Weiss, and Robert West. Political polarization in online news consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 152–162, 2021.

[Garrett *et al.*, 2014] R Kelly Garrett, Shira Dvir Gvirsman, Benjamin K Johnson, Yariv Tsfati, Rachel Neo, and Aysenur Dal. Implications of pro-and counterattitudinal information exposure for affective polarization. *Human communication research*, 40(3) :309–332, 2014.

[Ge *et al.*, 2010] Mouzhi Ge, Carla Delgado, and Dietmar Jannach. Beyond accuracy : Evaluating recommender systems by coverage and serendipity. In *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, pages 257–260, Barcelona, Spain, 01 2010. ACM.

[Geschke *et al.*, 2019] Daniel Geschke, Jan Lorenz, and Peter Holtz. The triple-filter bubble : Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology*, 58(1) :129–149, 2019.

[Giunchiglia *et al.*, 2021] Fausto Giunchiglia, Jahna Otterbacher, Styliani Kleanthous, Khuyagbaatar Batsuren, Veronika Bogin, Tsvi Kuflik, and Avital Shulner Tal. Towards algorithmic transparency : A diversity perspective. *arXiv preprint arXiv :2104.05658*, 2021.

[Goldstein, 1984] Jan Goldstein. "moral contagion" : a professional ideology of medicine and psychiatry in eighteenth-and nineteenth-century france. *Professions and the French State*, 1 :700–1900, 1984.

[Golman *et al.*, 2017] Russell Golman, David Hagmann, and George Loewenstein. Information avoidance. *Journal of economic literature*, 55(1) :96–135, 2017.

[Guerra *et al.*, 2013] Pedro Guerra, Wagner Meira Jr, Claire Cardie, and Robert Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 215–224, 2013.

[Gulla *et al.*, 2017] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. The adressa dataset for news recommendation. In *Proceedings of the international conference on web intelligence*, pages 1042–1048, Leipzig Germany, 2017. ACM.

[Gunawardana *et al.*, 2012] Asela Gunawardana, Guy Shani, and Sivan Yogev. Evaluating recommender systems. In *Recommender systems handbook*, pages 547–601. Springer, 2012.

[Guo *et al.*, 2021] Yanzhu Guo, Virgile Rennard, Christos Xypolopoulos, and Michalis Vazirgiannis. BERTweetFR : Domain adaptation of pre-trained language models for French tweets. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 445–450, Online, November 2021. Association for Computational Linguistics.

[Haim *et al.*, 2018] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. Burst of the filter bubble ? effects of personalization on the diversity of google news. *Digital journalism*, 6(3) :330–343, 2018.

[Harambam *et al.*, 2019] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. Designing for the better by taking users into account : A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM conference on recommender systems*, pages 69–77, 2019.

[Hastie, 2017] Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge, 2017.

[He *et al.*, 2023] Zhicheng He, Weiwen Liu, Wei Guo, Jiarui Qin, Yingxue Zhang, Yaochen Hu, and Ruiming Tang. A survey on user behavior modeling in recommender systems. *arXiv preprint arXiv :2302.11087*, 2023.

[Heinrich *et al.*, 2021] Bernd Heinrich, Marcus Hopf, Daniel Lohninger, Alexander Schiller, and Michael Szubartowicz. Data quality in recommender systems : the impact of completeness of item content data on prediction accuracy of recommender systems. *Electronic Markets*, 31 :389–409, 2021.

[Heitz *et al.*, 2022] Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. Benefits of diverse news recommendations for democracy : A user study. *Digital Journalism*, pages 1–21, 2022.

[Helberger *et al.*, 2018] Natali Helberger, Kari Karppinen, and Lucia D'acunto. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2) :191–207, 2018.

[Helberger, 2021] Natali Helberger. On the democratic role of news recommenders. In *Algorithms, Automation, and News*, pages 14–33. Routledge, 2021.

[Heltzel and Laurin, 2020] Gordon Heltzel and Kristin Laurin. Polarization in america : Two possible futures. *Current opinion in behavioral sciences*, 34 :179–184, 2020.

[Herlocker *et al.*, 2004] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1) :5–53, 2004.

[Hetherington, 2001] Marc J Hetherington. Resurgent mass partisanship : The role of elite polarization. *American political science review*, 95(3) :619–631, 2001.

[Himeur *et al.*, 2022] Yassine Himeur, Shahab Saquib Sohail, Faycal Bensaali, Abbes Amira, and Mamoun Alazab. Latest trends of security and privacy in recommender systems : a comprehensive review and future perspectives. *Computers & Security*, 118 :102746, 2022.

[Holland and Miller, 1991] John H Holland and John H Miller. Artificial adaptive agents in economic theory. *The American economic review*, 81(2) :365–370, 1991.

[Hu and Pu, 2011] Rong Hu and Pearl Pu. Helping users perceive recommendation diversity. In *Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), held in conjunction with RecSys'11*, pages 43–50, Chicago, USA, 2011. ACM.

[Huckfeldt, 1986] R Robert Huckfeldt. Politics in context : Assimilation and conflict in urban neighborhoodr new, 1986.

[Hurley, 2013] Neil J Hurley. Personalised ranking with diversity. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 379–382, 2013.

[Hyun and Moon, 2016] Ki Deuk Hyun and Soo Jung Moon. Agenda setting in the partisan tv news context : Attribute agenda setting and polarized evaluation of presidential candidates among viewers of nbc, cnn, and fox news. *Journalism & Mass Communication Quarterly*, 93(3) :509–529, 2016.

[Iyengar *et al.*, 2019] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22 :129–146, 2019.

[Jambor and Wang, 2010] Tamas Jambor and Jun Wang. Optimizing multiple objectives in collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 55–62, 2010.

[Jannach and Abdollahpouri, 2023] Dietmar Jannach and Himan Abdollahpouri. A survey on multi-objective recommender systems. *Frontiers in big Data*, 6 :1157899, 2023.

[Jannach and Adomavicius, 2016] Dietmar Jannach and Gediminas Adomavicius. Recommendations with a purpose. In *Proceedings of the 10th ACM conference on recommender systems*, pages 7–10, 2016.

[Jannach and Jugovac, 2019] Dietmar Jannach and Michael Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4) :1–23, 2019.

[Jannach and Zanker, 2012] Dietmar Jannach and Markus Zanker. Value and impact of recommender systems. In *Recommender systems handbook*, pages 519–546. Springer, 2012.

[Jannach *et al.*, 2020] Dietmar Jannach, Gabriel de Souza P. Moreira, and Even Oldridge. Why are deep learning models not consistently winning recommender systems competitions yet ? a position paper. In *Proceedings of the Recommender Systems Challenge 2020*, pages 44–49. 2020.

[Järvelin and Kekäläinen, 2002] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4) :422–446, 2002.

[Jasser *et al.*, 2022] Jasser Jasser, Ivan Garibay, Steve Scheinert, and Alexander V Mantzaris. Controversial information spreads faster and further than non-controversial information in reddit. *Journal of Computational Social Science*, 5(1) :111–122, 2022.

[Jawaheer *et al.*, 2014] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. Modeling user preferences in recommender systems : A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 4(2) :1–26, 2014.

[Ji *et al.*, 2016] Youchun Ji, Wenxing Hong, Yali Shangguan, Huan Wang, and Jing Ma. Regularized singular value decomposition in news recommendation system. In *2016 11th International Conference on Computer Science & Education (ICCSE)*, pages 621–626, Cambridge, UK, 2016. IEEE.

[Jung *et al.*, 2019] Jiin Jung, Patrick Grim, Daniel J Singer, Aaron Bramson, William J Berger, Bennett Holman, and Karen Kovaka. A multidisciplinary understanding of polarization. *American Psychologist*, 74(3) :301, 2019.

[Kalla and Smith, 2023] Dinesh Kalla and Nathan Smith. Study and analysis of chat gpt and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, 8(3) :827–833, 2023.

[Kaminskas and Bridge, 2016] Marius Kaminskas and Derek Bridge. Diversity, serendipity, novelty, and coverage : a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1) :1–42, 2016.

[Karimi *et al.*, 2018] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. News recommender systems–survey and roads ahead. *Information Processing & Management*, 54(6) :1203–1227, 2018.

[Katz *et al.*, 2017] Elihu Katz, Paul F Lazarsfeld, and Elmo Roper. *Personal influence : The part played by people in the flow of mass communications*. Routledge, 2017.

[Kaufman *et al.*, 2022] Miron Kaufman, Sanda Kaufman, and Hung T Diep. Statistical mechanics of political polarization. *Entropy*, 24(9) :1262, 2022.

[Kaya and Bridge, 2019a] Mesut Kaya and Derek Bridge. A comparison of calibrated and intent-aware recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 151–159, 2019.

[Kaya and Bridge, 2019b] Mesut Kaya and Derek Bridge. Subprofile-aware diversification of recommendations. *User Modeling and User-Adapted Interaction*, 29 :661–700, 2019.

[Kille *et al.*, 2013] Benjamin Kille, Frank Hopfgartner, Torben Brodt, and Tobias Heintz. The plista dataset. In *Proceedings of the 2013 international news recommender systems workshop and challenge*, pages 16–23, Kowloon Hong Kong, 2013. ACM.

[Kim, 2015] Yonghwan Kim. Does disagreement mitigate polarization ? how selective exposure and disagreement affect political polarization. *Journalism & Mass Communication Quarterly*, 92(4) :915–937, 2015.

[Kim, 2019] Yonghwan Kim. How cross-cutting news exposure relates to candidate issue stance knowledge, political polarization, and participation : The moderating role of political sophistication. *International Journal of Public Opinion Research*, 31(4) :626–648, 2019.

[Klayman, 1995] Joshua Klayman. Varieties of confirmation bias. *Psychology of learning and motivation*, 32 :385–418, 1995.

[Knijnenburg and Willemsen, 2015] Bart P Knijnenburg and Martijn C Willemsen. Evaluating recommender systems with user experiments. In *Recommender systems handbook*, pages 309–352. Springer, 2015.

[Knobloch-Westerwick *et al.*, 2015] Silvia Knobloch-Westerwick, Cornelia Mothes, Benjamin K Johnson, Axel Westerwick, and Wolfgang Donsbach. Political online information searching in germany and the united states : Confirmation bias, source credibility, and attitude impacts. *Journal of Communication*, 65(3) :489–511, 2015.

[Ko *et al.*, 2022] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems : recommendation models, techniques, and application fields. *Electronics*, 11(1) :141, 2022.

[Konstan *et al.*, 1998] Joseph A Konstan, John Riedl, Al Borchers, and Jonathan L Herlocker. Recommender systems : A grouplens perspective. In *Recommender Systems : Papers from the 1998 Workshop (AAAI Technical Report WS-98-08)*, pages 60–64. AAAI Press Menlo Park, 1998.

[Koudenburg *et al.*, 2021] Namkje Koudenburg, Henk AL Kiers, and Yoshihisa Kashima. A new opinion polarization index developed by integrating expert judgments. *Frontiers in psychology*, 12 :738258, 2021.

[Kriesi, 2017] Hanspeter Kriesi. The populist challenge. In *The Role of Parties in Twenty-First Century Politics*, pages 131–148. Routledge, 2017.

[Kubin and Von Sikorski, 2021] Emily Kubin and Christian Von Sikorski. The role of (social) media in political polarization : a systematic review. *Annals of the International Communication Association*, 45(3) :188–206, 2021.

[Kunaver and Požrl, 2017] Matevž Kunaver and Tomaž Požrl. Diversity in recommender systems–a survey. *Knowledge-based systems*, 123 :154–162, 2017.

[Lathia *et al.*, 2010] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, 2010.

[Lazarsfeld *et al.*, 1968] Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. *The people's choice : How the voter makes up his mind in a presidential campaign.* Columbia University Press, 1968.

[Lazer *et al.*, 2018] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380) :1094–1096, 2018.

[Leonhardt *et al.*, 2018] Jurek Leonhardt, Avishek Anand, and Megha Khosla. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*, pages 101–102, 2018.

[Levin *et al.*, 2021] Simon A Levin, Helen V Milner, and Charles Perrings. The dynamics of political polarization, 2021.

[Levy, 2021] Ro'ee Levy. Social media, news consumption, and polarization : Evidence from a field experiment. *American economic review*, 111(3) :831–870, 2021.

[L'Huillier *et al.*, 2014] Amaury L'Huillier, Sylvain Castagnos, and Anne Boyer. Understanding usages by modeling diversity over time. In *In extended proceedings of the 22nd Conference on User Modelling, Adaptation and Personalization (UMAP 2014)*, Aalborg, Denmark, 07 2014. ACM.

[Li *et al.*, 2023] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation : Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5) :1–48, 2023.

[Lika *et al.*, 2014] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. Facing the cold start problem in recommender systems. *Expert systems with applications*, 41(4) :2065–2073, 2014.

[Likas *et al.*, 2003] Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, pages 451–461, 2003.

[Lops *et al.*, 2011] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. Content-based recommender systems : State of the art and trends. *Recommender systems handbook*, pages 73–105, 2011.

[Lops *et al.*, 2023] Pasquale Lops, Marco Polignano, Cataldo Musto, Antonio Silletti, and Giovanni Semeraro. Clayrs : An end-to-end framework for reproducible knowledge-aware recommender systems. *Information Systems*, 119 :102273, 2023.

[Lorenz-Spreen *et al.*, 2020] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R Sunstein, and Ralph Hertwig. How behavioural sciences can promote truth, autonomy and democratic discourse online. *Nature human behaviour*, 4(11) :1102–1109, 2020.

[Lu and Tintarev, 2018] Feng Lu and Nava Tintarev. A diversity adjusting strategy with personality for music recommendation. In *IntRS@ RecSys*, pages 7–14, 2018.

[Lubos *et al.*, 2023] Sebastian Lubos, Alexander Felfernig, and Markus Tautschnig. An overview of video recommender systems : state-of-the-art and research issues. *Frontiers in big Data*, 6, 2023.

[Lunardi *et al.*, 2020] Gabriel Machado Lunardi, Guilherme Medeiros Machado, Vinicius Maran, and José Palazzo M. de Oliveira. A metric for filter bubble measurement in recommender algorithms considering the news domain. *Applied Soft Computing*, 97 :106771, 2020.

[Lund and Zhong, 2018] Campbell Lund and Shirui Zhong. The impact of tiktok's engagement algorithm on political polarization. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym'XX) Vol*, volume 1, pages 1–8, 2018.

[Luostarinen and Kohonen, 2013] Tapio Luostarinen and Oskar Kohonen. Using topic models in content-based news recommender systems. In *Proceedings of the 19th Nordic conference of computational linguistics (NODALIDA 2013)*, pages 239–251, 2013.

[Ma *et al.*, 2016] Hao Ma, Xueqing Liu, and Zhihong Shen. User fatigue in online news recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1363–1372, 2016.

[Macy *et al.*, 2021] Michael W Macy, Manqing Ma, Daniel R Tabin, Jianxi Gao, and Boleslaw K Szymanski. Polarization and tipping points. *Proceedings of the National Academy of Sciences*, 118(50) :e2102144118, 2021.

[Maksai *et al.*, 2015] Andrii Maksai, Florent Garcin, and Boi Faltings. Predicting online performance of news recommender systems through richer evaluation metrics. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 179–186, Vienna, Austria, 2015. ACM.

[Manning *et al.*, 2008] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge university press, 2008.

[Marozzo and Bessi, 2018] Fabrizio Marozzo and Alessandro Bessi. Analyzing polarization of social media users and news sites during political campaigns. *Social Network Analysis and Mining*, 8 :1–13, 2018.

[McCombs and Shaw, 1972] Maxwell E McCombs and Donald L Shaw. The agenda-setting function of mass media. *Public opinion quarterly*, 36(2) :176–187, 1972.

[McCoy *et al.*, 2018] Jennifer McCoy, Tahmina Rahman, and Murat Somer. Polarization and the global crisis of democracy : Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62(1) :16–42, 2018.

[McCrae and John, 1992] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2) :175–215, 1992.

[McInnes *et al.*, 2018] Leland McInnes, John Healy, and James Melville. Umap : Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv :1802.03426*, 2018.

[McNee *et al.*, 2006] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough : how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101, 2006.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather : Homophily in social networks. *Annual review of sociology*, 27(1) :415–444, 2001.

[Milačić, 2021] Filip Milačić. The negative impact of polarization on democracy, 2021.

[Milano *et al.*, 2020] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. Recommender systems and their ethical challenges. *Ai & Society*, 35 :957–967, 2020.

[Miller and Page, 2008] John Howard Miller and Scott E Page. Complex adaptive systems : an introduction to computational models of social life. *(No Title)*, 2008.

[Morales *et al.*, 2015] Alfredo Jose Morales, Javier Borondo, Juan Carlos Losada, and Rosa M Benito. Measuring political polarization : Twitter shows the two sides of venezuela. *Chaos : An Interdisciplinary Journal of Nonlinear Science*, 25(3) :033114, 2015.

[Moulavi *et al.*, 2014] Davoud Moulavi, Pablo A Jaskowiak, Ricardo JGB Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 839–847. SIAM, 2014.

[Nadeau and Sekine, 2007] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26, 2007.

[Nettleton, 2013] David F Nettleton. Data mining of social networks represented as graphs. *Computer Science Review*, 7 :1–34, 2013.

[Newman, 2006] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23) :8577–8582, 2006.

[Nguyen, 2020] C Thi Nguyen. Echo chambers and epistemic bubbles. *Episteme*, 17(2) :141–161, 2020.

[Nordgren, 2022] Anders Nordgren. Artificial intelligence and climate change : ethical issues. *Journal of Information, Communication and Ethics in Society*, 21(1) :1–15, 2022.

[Norris, 2003] Pippa Norris. Preaching to the converted ? pluralism, participation and party websites. *Party politics*, 9(1) :21–45, 2003.

[Papagelis *et al.*, 2005] Manos Papagelis, Ioannis Rousidis, Dimitris Plexousakis, and Elias Theoharopoulos. Incremental collaborative filtering for highly-scalable recommendation algorithms. In *Foundations of Intelligent Systems : 15th International Symposium, ISMIS 2005, Saratoga Springs, NY, USA, May 25-28, 2005. Proceedings 15*, pages 553–561. Springer, 2005.

[Pariser, 2011] Eli Pariser. *The filter bubble : What the Internet is hiding from you.* penguin UK, 2011.

[Park and Tuzhilin, 2008] Yoon-Joo Park and Alexander Tuzhilin. The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 11–18, 2008.

[Parra and Sahebi, 2013] Denis Parra and Shaghayegh Sahebi. Recommender systems : Sources of knowledge and evaluation metrics. In *Advanced Techniques in Web Intelligence-2 : Web User Browsing Behaviour and Preference Analysis*, pages 149–175. Springer, 2013.

[Patel *et al.*, 2017] Bansari Patel, Palak Desai, and Urvi Panchal. Methods of recommender system : A review. In *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)*, pages 1–4. IEEE, 2017.

[Patino, 2019] Bruno Patino. *La civilisation du poisson rouge : petit traité sur le marché de l'attention.* Grasset, 2019.

[Pazzani and Billsus, 2007] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web : methods and strategies of web personalization*, pages 325–341. Springer, 2007.

[Peralta *et al.*, 2024] Antonio F Peralta, Pedro Ramaciotti, János Kertész, and Gerardo Iñiguez. Multidimensional political polarization in online social networks. *Physical Review Research*, 6(1) :013170, 2024.

[Pereira *et al.*, 2018] Fabíola SF Pereira, João Gama, Sandra de Amo, and Gina MB Oliveira. On analyzing user preference dynamics with temporal social networks. *Machine Learning*, 107 :1745–1773, 2018.

[Perra and Rocha, 2019] Nicola Perra and Luis EC Rocha. Modelling opinion dynamics in the age of algorithmic personalisation. *Scientific reports*, 9(1) :7261, 2019.

[Perrissol and Somat, 2009] Stéphane Perrissol and Alain Somat. L'exposition sélective : bilan et perspectives. *L'année psychologique*, 109(3) :551–581, 2009.

[Phillips *et al.*, 2023] Samantha C Phillips, Joshua Uyheng, and Kathleen M Carley. A high-dimensional approach to measuring online polarization. *Journal of Computational Social Science*, 6(2) :1147–1178, 2023.

[Primario *et al.*, 2017] Simonetta Primario, Dario Borrelli, Luca Iandoli, Giuseppe Zollo, and Carlo Lipizzi. Measuring polarization in twitter enabled in online political conversation : The case of 2016 us presidential election. In *2017 IEEE international conference on information reuse and integration (IRI)*, pages 607–613. IEEE, 2017.

[Prior, 2013] Markus Prior. Media and political polarization. *Annual review of political science*, 16 :101–127, 2013.

[Rabiu *et al.*, 2020] Idris Rabiu, Naomie Salim, Aminu Da'u, and Akram Osman. Recommender system based on temporal models : a systematic review. *Applied Sciences*, 10(7) :2204, 2020.

[Raza and Ding, 2022] Shaina Raza and Chen Ding. News recommender system : a review of recent progress, challenges, and opportunities. *Artificial Intelligence Review*, pages 1–52, 2022.

[Raza, 2021] Shaina Raza. A news recommender system considering temporal dynamics and diversity. *arXiv preprint arXiv :2103.12537*, 2021.

[Resnick and Varian, 1997] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3) :56–58, 1997.

[Resnick *et al.*, 1994] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens : An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, pages 175–186, 1994.

[Ricci *et al.*, 2011] F. Ricci, L. Rokach, B. Shapira, and P. Kantor. *Recommender systems handbook*. Springer, Berlin, Germany, 2011.

[Ricci *et al.*, 2021] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems : Techniques, applications, and challenges. *Recommender systems handbook*, pages 1–35, 2021.

[Rousseeuw, 1987] Peter J. Rousseeuw. Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, pages 53–65, 1987.

[Roy and Dutta, 2022] Deepjyoti Roy and Mala Dutta. A systematic review and research perspective on recommender systems. *Journal of Big Data*, 9(1) :59, 2022.

[Russell Neuman *et al.*, 2014] W Russell Neuman, Lauren Guggenheim, Set al Mo Jang, and Soo Young Bae. The dynamics of public attention : Agenda-setting theory meets big data. *Journal of communication*, 64(2) :193–214, 2014.

[Russell, 2021] Annelise Russell. Minority opposition and asymmetric parties ? senators' partisan rhetoric on twitter. *Political Research Quarterly*, 74(3) :615–627, 2021.

[Said *et al.*, 2013] Alan Said, Ben Fields, Brijnesh J Jain, and Sahin Albayrak. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1399–1408, 2013.

[Sakoe and Chiba, 1978] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1) :43–49, 1978.

[Santos *et al.*, 2010] Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th international conference on World wide web*, pages 881–890, 2010.

[Sanz-Cruzado and Castells, 2018] Javier Sanz-Cruzado and Pablo Castells. Enhancing structural diversity in social networks by recommending weak ties. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, page 233–241, Vancouver, British Columbia, Canada, 2018. ACM.

[Sarwar *et al.*, 2000] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*, pages 158–167, 2000.

[Schelenz, 2021] Laura Schelenz. Diversity-aware recommendations for social justice ? exploring user diversity and fairness in recommender systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 404–410, 2021.

[Scheufele, 1999] Dietram A Scheufele. Framing as a theory of media effects. *Journal of communication*, 49(1) :103–122, 1999.

[Schmidt *et al.*, 2018] Ana Lucía Schmidt, Fabiana Zollo, Antonio Scala, Cornelia Betsch, and Walter Quattrociocchi. Polarization of the vaccination debate on facebook. *Vaccine*, 36(25) :3606–3612, 2018.

[Schmitt *et al.*, 2018] Josephine B Schmitt, Christina A Debbelt, and Frank M Schneider. Too much information ? predictors of information overload in the context of online news exposure. *Information, Communication & Society*, 21(8) :1151–1167, 2018.

[Seymen *et al.*, 2021] Sinan Seymen, Himan Abdollahpouri, and Edward C Malthouse. A constrained optimization approach for calibrated recommendations. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 607–612, 2021.

[Shannon, 1948] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3) :379–423, 1948.

[Shi *et al.*, 2012] Yue Shi, Xiaoxue Zhao, Jun Wang, Martha Larson, and Alan Hanjalic. Adaptive diversification of recommendation results via latent factor portfolio. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 175–184, 2012.

[Simon, 1957] Herbert A Simon. Models of man ; social and rational. 1957.

[Sîrbu *et al.*, 2017] Alina Sîrbu, Vittorio Loreto, Vito DP Servedio, and Francesca Tria. Opinion dynamics : models, extensions and external effects. *Participatory sensing, opinions and collective awareness*, pages 363–401, 2017.

[Smyth and McClave, 2001] Barry Smyth and Paul McClave. Similarity vs. diversity. In David W. Aha and Ian Watson, editors, *Case-Based Reasoning Research and Development*, pages 347–361, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[Sparck Jones, 1972] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1) :11–21, 1972.

[Spohr, 2017] Dominic Spohr. Fake news and ideological polarization : Filter bubbles and selective exposure on social media. *Business information review*, 34(3) :150–160, 2017.

[Steck, 2018] Harald Steck. Calibrated recommendations. In *Proceedings of the 12th ACM conference on recommender systems*, pages 154–162, 2018.

[Stray, 2021] Jonathan Stray. Designing recommender systems to depolarize. *arXiv preprint arXiv :2107.04953*, 2021.

[Su *et al.*, 2013] Ruilong Su, Li'Ang Yin, Kailong Chen, and Yong Yu. Set-oriented personalized ranking for diversified top-n recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 415–418, 2013.

[Sunstein, 1999] Cass R Sunstein. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper*, (91), 1999.

[Sunstein, 2009] Cass R. Sunstein. *Going to extremes : how like minds unite and divide*. Oxford University Press, 2009.

[Tardelli *et al.*, 2023] Serena Tardelli, Leonardo Nizzoli, Maurizio Tesconi, Mauro Conti, Preslav Nakov, Giovanni Da San Martino, and Stefano Cresci. Temporal dynamics of coordinated online behavior : Stability, archetypes, and influence. *arXiv preprint arXiv :2301.06774*, 2023.

[Thorat *et al.*, 2015] Poonam B Thorat, Rajeshwari M Goudar, and Sunita Barve. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, 110(4) :31–36, 2015.

[Tintarev *et al.*, 2018] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. Same, same, but different : algorithmic diversification of viewpoints in news. In *Adjunct publication of the 26th conference on user modeling, adaptation and personalization*, pages 7–13, 2018.

[Tokita *et al.*, 2021] Christopher K Tokita, Andrew M Guess, and Corina E Tarnita. Polarized information ecosystems can reorganize social networks via information cascades. *Proceedings of the National Academy of Sciences*, 118(50) :e2102147118, 2021.

[Torrey and Shavlik, 2010] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends : algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[Trattner *et al.*, 2022] Christoph Trattner, Dietmar Jannach, Enrico Motta, Irene Costera Meijer, Nicholas Diakopoulos, Mehdi Elahi, Andreas L Opdahl, Bjørnar Tessem, Njål Borch, Morten Fjeld, et al. Responsible media technology and ai : challenges and research directions. *AI and Ethics*, 2(4) :585–594, 2022.

[Treuillier *et al.*, 2022] Celina Treuillier, Sylvain Castagnos, Evan Dufraisse, and Armelle Brun. Being diverse is not enough : Rethinking diversity evaluation to meet challenges of news recommender systems. In *Fairness in User Modeling, Adaptation and Personalization (FairUMAP 2022)*, 2022.

[Treuillier *et al.*, 2024a] Celina Treuillier, Sylvain Castagnos, Christèle Lagier, and Armelle Brun. Gaining a better understanding of online polarization by approaching it as a dynamic process. *Scientific Reports*, 14(1) :8702, 2024.

[Treuillier *et al.*, 2024b] Celina Treuillier, Özlem Özgöbek, Sylvain Castagnos, and Armelle Brun. Beyond trade-offs : Unveiling fairness-constrained diversity in news recommender systems. In *Proceedings of the 32st ACM Conference on User Modeling, Adaptation and Personalization*, 2024.

[Turenne, 2018] Nicolas Turenne. The rumour spectrum. *PloS one*, 13(1) :e0189080, 2018.

[Valensise *et al.*, 2023] Carlo M Valensise, Matteo Cinelli, and Walter Quattrociocchi. The drivers of online polarization : fitting models to data. *Information Sciences*, page 119152, 2023.

[Valenzuela *et al.*, 2021] Sebastián Valenzuela, Ingrid Bachmann, and Matías Bargsted. The personal is the political ? what do whatsapp users share and how it matters for news knowledge, polarization and participation in chile. *Digital journalism*, 9(2) :155–175, 2021.

[Van Bavel *et al.*, 2021] Jay J Van Bavel, Steve Rathje, Elizabeth Harris, Claire Robertson, and Anni Sternisko. How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11) :913–916, 2021.

[Van Stekelenburg, 2014] Jacquelien Van Stekelenburg. Going all the way : Politicizing, polarizing, and radicalizing identity offline and online. *Sociology Compass*, 8(5) :540–555, 2014.

[Vargas and Castells, 2011] Saul Vargas and Pablo Castells. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys'11*, pages 109–116, Chicago, USA, 2011. ACM.

[Vargas and Castells, 2013] Saúl Vargas and Pablo Castells. Exploiting the diversity of user preferences for recommendation. In *Proceedings of the 10th conference on open research areas in information retrieval*, pages 129–136. Citeseer, 2013.

[Vargas *et al.*, 2011] Saul Vargas, Pablo Castells, and David Vallet. Intent-oriented diversity in recommender systems. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1211–1212, 2011.

[Vargas *et al.*, 2012] Saúl Vargas, Pablo Castells, and David Vallet. Explicit relevance models in intent-oriented information retrieval diversification. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 75–84, 2012.

[Vargas, 2015] Saúl Vargas. *Novelty and diversity enhancement and evaluation in Recommender Systems*. PhD thesis, Universidad Autónoma de Madrid, 2015.

[Viana and Soares, 2016] Paula Viana and Márcio Soares. A hybrid recommendation system for news in a mobile environment. In *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*, pages 1–9, 2016.

[Vrijenhoek, 2023] Sanne Vrijenhoek. Do you mind ? reflections on the mind dataset for research on diversity in news recommendations. In *International Workshop on Algorithmic Bias in Search and Recommendation*, pages 147–154. Springer, 2023.

[Walker and Matsa, 2021] Mason Walker and Katerina Eva Matsa. News consumption across social media in 2021. 2021.

[Waller and Anderson, 2021] Isaac Waller and Ashton Anderson. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888) :264–268, 2021.

[Wang *et al.*, 2018] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. Dkn : Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*, pages 1835–1844, 2018.

[Wang *et al.*, 2023] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems*, 41(3) :1–43, 2023.

[Wu *et al.*, 2018] Wen Wu, Li Chen, and Yu Zhao. Personalizing recommendation diversity based on user personality. *User Modeling and User-Adapted Interaction*, 28(3) :237–276, 2018.

[Wu *et al.*, 2020a] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Sentirec : Sentiment diversity-aware neural news recommendation. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*, pages 44–53, 2020.

[Wu *et al.*, 2020b] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. Mind : A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online, 2020. ACL.

[Wu *et al.*, 2023] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. Personalized news recommendation : Methods and challenges. *ACM Transactions on Information Systems*, 41(1) :1–50, 2023.

[Yarchi *et al.*, 2021] Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. Political polarization on the digital sphere : A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38(1-2) :98–139, 2021.

[Yunanda *et al.*, 2022] Gisela Yunanda, Dade Nurjanah, and Selly Meliana. Recommendation system from microsoft news data using tf-idf and cosine similarity methods. *Building of Informatics, Technology and Science (BITS)*, 4(1) :277–284, 2022.

[Zahra *et al.*, 2015] Sobia Zahra, Mustansar Ali Ghazanfar, Asra Khalid, Muhammad Awais Azam, Usman Naeem, and Adam Prugel-Bennett. Novel centroid selection approaches for kmeans-clustering based recommender systems. *Information sciences*, 320 :156–189, 2015.

[Zangerle and Bauer, 2022] Eva Zangerle and Christine Bauer. Evaluating recommender systems : survey and framework. *ACM computing surveys*, 55(8) :1–38, 2022.

[Zhai *et al.*, 2015] ChengXiang Zhai, William W Cohen, and John Lafferty. Beyond independent relevance : methods and evaluation metrics for subtopic retrieval. In *Acm sigir forum*, volume 49, pages 2–9. ACM New York, NY, USA, 2015.

[Zhang, 2013] Liang Zhang. The definition of novelty in recommendation system. *Journal of Engineering Science and Technology Review*, 6(3) :141–145, 2013.

[Zhou *et al.*, 2010] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10) :4511–4515, 2010.

[Ziegler *et al.*, 2005] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

[Zuiderveen Borgesius *et al.*, 2016] Frederik Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. Should we worry about filter bubbles ? *Internet Policy Review. Journal on Internet Regulation*, 5(1), 2016.

Résumé

La polarisation est un phénomène complexe exacerbé par les médias sociaux, auquel les systèmes de recommandation contribuent en limitant l'exposition à des opinions diverses. Pour contrer cet enjeu sociétal et favoriser un débat démocratique sain, il est crucial de développer des stratégies de recommandation favorisant une dépolarisation personnalisée. Une compréhension du phénomène de polarisation est néanmoins indispensable au développement de telles stratégies.

Dans une première partie, cette thèse propose une modélisation individuelle et multi-factorielle du phénomène de polarisation. Un modèle générique et paramétrable, reposant sur la métrique GRAIL (GeneRalized AddItive poLarization), permet de distinguer des classes de comportement et contribue à une meilleure compréhension de la polarisation. Une modélisation temporelle abordant la polarisation comme un processus dynamique permet ensuite d'identifier des périodes de polarisation, influencées par la maturité des débats et des événements contextuels. Dans une seconde partie, l'approche de diversification personnalisée ADF (Accuracy-Diversity-Fairness) est proposée. Les approches de la littérature peinent à maîtriser la nature de la diversité apportée, et peuvent artificiellement orienter les opinions des utilisateurs. L'approche ADF propose une optimisation tri-objectif permettant de fournir des recommandations répondant aux attentes des utilisateurs, tout en les exposant à un contenu plus diversifié sans influencer leurs opinions.

Ces travaux participent à une modélisation plus fine du phénomène de polarisation, permettant d'envisager développement de stratégies dépolarisantes efficaces et éthiques. Ce travail met en lumière le rôle central que joue l'IA pour répondre aux enjeux sociétaux contemporains.

Mots-clés: I.A., Systèmes de recommandation, Diversité, Équité, Polarisation, Médias sociaux

Abstract

Polarization is a complex phenomenon exacerbated by social media, where recommender systems limit exposure to diverse opinions. To counter this societal challenge and promote healthy democratic debate, it is crucial to develop recommendation strategies that encourage personalized depolarization. Understanding polarization is essential for developing such strategies.

In the first part, this thesis proposes an individual and multi-factorial model of polarization. A generic, configurable model, based on the GRAIL (GeneRalized AddItive poLarization) metric, allows to discriminate between behavioral classes and enhances understanding of polarization. Additionally, a temporal model approaching polarization as a dynamic process highlights periods of polarization, influenced by the maturity of debates and contextual events. In the second part, a personalized diversification approach, ADF (Accuracy-Diversity-Fairness), is proposed. Existing approaches struggle to control the nature of the diversity provided and can artificially orient users' opinions. ADF proposes a tri-objective optimization to provide recommendations that meet users' needs, while exposing them to more diverse content without influencing their opinions.

This work contributes to a more detailed modeling of polarization, enabling the development of effective and ethical depolarization strategies. It highlights the pivotal role of AI in addressing contemporary societal challenges.

Keywords: A.I., Recommender Systems, Diversity, Equity, Polarization, Social Media