Modèles thématiques pour la découverte non supervisée de points de vue sur le Web

Thibaut Thonet

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *23/11/2017* **par :**

Thibaut THONET

Modèles thématiques pour la découverte non supervisée de points de vue sur le Web

---

JURY

Nathalie AUSSENAC-GILLES	DR au CNRS, Université Toulouse 3	Présidente du jury
Patrick GALLINARI	Professeur, Université Pierre et Marie Curie	Rapporteur
Éric GAUSSIER	Professeur, Université Grenoble-Alpes	Rapporteur
Julien VELCIN	MCF-HDR, Université Lyon 2	Examinateur
Guillaume CABANAC	MCF-HDR, Université Toulouse 3	Directeur de thèse
Karen PINEL-SAUVAGNAT	MCF, Université Toulouse 3	Co-encadrante
Mohand BOUGHANEM	Professeur, Université Toulouse 3	Co-encadrant

---

École doctorale et spécialité :
MITT : Image, Information, Hypermédia

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :
Guillaume CABANAC, *Karen PINEL-SAUVAGNAT* et *Mohand BOUGHANEM*

Rapporteurs :
Patrick GALLINARI et *Éric GAUSSIER*

Remerciements

Cette thèse n'aurait pu aboutir sans l'aide bienveillante et le soutien indéfectible que m'ont témoignés les nombreuses personnes côtoyées durant ces quelques années à l'IRIT et au delà. Mes premiers remerciements vont tout naturellement à mes encadrants de thèse, Monsieur Guillaume Cabanac, Madame Karen Pinel-Sauvagnat et Monsieur Mohand Boughanem. Je ne saurais trouver les mots pour exprimer la gratitude que j'éprouve à leur égard pour leur patience, leur sollicitude et leurs encouragements de tous les instants, qui m'ont permis de continuer à avancer même lorsque le moral manquait. Plus que des encadrants, je peux sans le moindre doute affirmer avoir trouver en eux des mentors, des exemples à suivre, qui m'ont transmis leur passion pour leur profession.

Je souhaite également exprimer toute ma reconnaissance aux rapporteurs de ma thèse, Monsieur Patrick Gallinari et Monsieur Éric Gaussier, qui m'ont fait l'honneur d'apporter leur précieuse expertise dans l'évaluation de mes travaux. Au delà de son rôle de rapporteur, je remercie par ailleurs Éric pour son aide dans la relecture de mes articles, pour nos multiples discussions ou correspondances passionnantes, ainsi que pour son invitation à effectuer un séminaire au laboratoire LIG de l'Université Grenoble Alpes – opportunité qui s'est avérée fort enrichissante pour moi.

Mes sincères remerciements vont de même à Madame Nathalie Aussenac-Gilles et Monsieur Julien Velcin, qui ont gracieusement accepté de siéger dans mon jury de thèse en tant qu'examinatrice et examinateur, respectivement. Je remercie par ailleurs Julien pour nos nombreux échanges sur l'évaluation des modèles thématiques et pour sa participation au groupe de travail CaRaThoVe, aux cotés de Monsieur Pierre Ratinaud et Monsieur Guillaume Cabanac – collaboration qui, j'en suis convaincu, ne tardera pas à porter ses fruits.

Cette thèse n'aurait eu la même saveur sans toutes les personnes que j'ai rencontrées et côtoyées à l'IRIT. Mes remerciements vont spécialement à celles et ceux avec qui j'ai partagé le bureau 406. J'ai une reconnaissance toute particulière pour Diep qui m'a accompagné durant toutes ces années de thèse et m'a toujours gratifié de son soutien sans faille. Plus généralement, je remercie les différents membres de l'équipe IRIS, passés ou présents, permanents ou doctorants.

Enfin, j'éprouve une immense gratitude pour ma famille et mes amis qui m'ont soutenu – voire supporté – durant ces années de thèse parfois éprouvantes mais toujours exaltantes. Je remercie en particulier mes parents pour leur enthousiasme, leur disponibilité, leur patience et leur amour inconditionnel, sans lesquels cette thèse n'aurait été possible.

Table des matières

Table des figures

Liste des tableaux

Publications

Articles de conférences internationales avec comité de lecture

1. **Thibaut Thonet**, Guillaume Cabanac, Mohand Boughanem, Karen Pinel-Sauvagnat (2017). Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, CIKM '17. *(À paraître.)*
2. **Thibaut Thonet**, Guillaume Cabanac, Mohand Boughanem, Karen Pinel-Sauvagnat (2016). VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery. In *Proceedings of the 38th European Conference on IR Research*, ECIR '16, pages 533–545.

Article de conférence nationale avec comité de lecture

1. **Thibaut Thonet**, Romain Deveaud, Iadh Ounis, Craig Macdonald (2015). Suggestion Contextuelle Composite. In *Actes de la 12ème Conférence en Recherche d'Information et Applications*, CORIA '15, pages 89–104.

Article de campagne d'évaluation internationale (sans comité de lecture)

1. Richard McCreadie, Romain Deveaud, M-Dyaa Albakour, Stuart Mackie, Nut Limsopatham, Craig Macdonald, Iadh Ounis, **Thibaut Thonet**, Bekir Taner Dinçer (2014). University of Glasgow at TREC 2014: Experiments with Terrier in Contextual Suggestion, Temporal Summarisation and Web Tracks. In *Proceedings of the 23rd Text Retrieval Conference*, TREC '14.

Introduction générale

1 Contexte

À une ère où le numérique est omniprésent et où le Web rythme nos vies quotidiennes, nous disposons de moyens auparavant inégalés par leur ampleur et leur rapidité pour nous informer, communiquer et partager du contenu. Cette révolution de possibilités est notamment le fruit de l'émergence de plateformes en ligne permettant à leurs utilisateurs de s'exprimer ou d'interagir. À travers ces plateformes, les internautes ont alors accès à un outil simple à prendre en main pour diffuser leurs *opinions*. On peut citer en exemple le site de commerce en ligne Amazon [1] – qui permet aux acheteurs de noter et critiquer les produits commandés – ou la plateforme de *microblogging* Twitter [2] – qui offre à ses utilisateurs la possibilité de s'exprimer publiquement sous la forme de courts messages de 140 caractères.

Loin d'être anodine, cette richesse d'opinions publiées sur le Web s'est rapidement avérée d'importance capitale pour une multitude d'acteurs économiques. En effet, nous ne sommes pour la plupart pas insensibles aux opinions partagées par les autres internautes, et en particulier dans le cas de l'achat en ligne. Un sondage du *Pew Research Center* [3] réalisé en 2016 va dans ce sens : 82 % des personnes (de nationalité américaine) interrogées ont déclaré consulter les notes et les critiques en ligne lorsqu'elles achètent un produit pour la première fois [Smith et Anderson, 2016]. La masse de données d'opinions disponible – et les promesses qu'elle offre – a alors motivé à la fois chercheurs universitaires et industriels à développer des systèmes de fouille de textes focalisés sur ces données subjectives. Ce domaine d'étude est aujourd'hui connu sous le nom de *fouille d'opinions* (*opinion mining*) ou d'*analyse de sentiments* (*sentiment analysis*) [Pang et Lee, 2008].

1.1 Fouille d'opinions sur le Web

La fouille d'opinions, développée à la fin des années 1990 [Hatzivassiloglou et McKeown, 1997], se positionne à l'intersection de plusieurs spécialités de l'informatique telles que le traitement automatique du langage naturel, la recherche d'information, la fouille de texte et l'apprentissage automatique. Elle est définie par Pang et Lee [2008] comme étant « le traitement informatique de l'opinion, du sentiment et de la subjectivité dans le texte ». Dans le faits, les travaux en fouille d'opinions se sont longtemps focalisés sur les opinions formulées dans les critiques en ligne, telles que celles trouvées sur Amazon. Ce type d'opinions est typiquement associé à une polarité positive ou négative (voire un intermédiaire entre ces deux extrêmes), indiquant l'avis de l'internaute vis-à-vis du produit acheté.

1. `https://www.amazon.fr/`
2. `https://twitter.com/`
3. `http://www.pewresearch.org/`

Les approches classiques de fouille d'opinions, traditionnellement basées sur des lexiques de mots d'opinions [Jo et Oh, 2011; Lin et He, 2009; Yu et Hatzivassiloglou, 2003], sont toutefois difficilement utilisables lorsque l'on souhaite étudier des opinions plus complexes telles que les opinions politiques. Par exemple, dans le contexte de l'élection présidentielle américaine de 2016, le fait qu'un utilisateur de médias sociaux emploie fréquemment un lexique positif ou négatif ne sera pas nécessairement révélateur de son soutien pour Donald Trump ou Hillary Clinton. Des techniques nouvelles sont alors requises pour aller au delà de la catégorisation des opinions en « positives » et « négatives ». Nous regroupons ici ces techniques sous le nom de *fouille de points de vue*, où la notion de point de vue généralise l'opinion au delà de son acception usuelle liée à la polarité (positive ou négative).

1.2 Vers la fouille de points de vue

Il est pertinent de se demander dans quelle mesure la fouille de points de vue – qui présente les défis inédits évoqués précédemment – est nécessaire et ce qu'elle peut apporter. Le besoin d'avancées en fouille de points de vue repose – en partie – sur l'importance qu'a revêtu la discussion politique dans les médias sociaux et sur le Web en général. Par exemple, un sondage du *Pew Research Center* [Duggan et Smith, 2016] indique qu'en 2016 environ un tiers des utilisateurs de médias sociaux de nationalité américaine commentent, participent à des discussions ou postent du contenu en rapport avec le gouvernement et la politique. Cet important volume de données d'opinions politiques semble alors susceptible de pouvoir complémenter (à défaut de supplanter [Kim *et al.*, 2014]) les opinions recueillies par les sondages classiques [O'Connor *et al.*, 2010].

Ce même rapport du *Pew Research Center* [Duggan et Smith, 2016] relate par ailleurs la propension des utilisateurs de médias sociaux à s'abonner aux personnalités politiques qui partagent le même point de vue : 66 % des utilisateurs s'abonnent à des personnalités politiques qui partagent le même point de vue, alors que seulement 31 % (respectivement, 3 %) s'abonnent à des personnalités politiques aux points de vue variés (respectivement, opposés aux leurs). Ainsi, les internautes auront plus tendance à être exposés à du contenu soutenant leur propre point de vue qu'à du contenu qui y est opposé. Ce phénomène récemment mis à jour est connu sous le nom de « chambre d'échos » [Sunstein, 2009] et de « bulle de filtres » [Pariser, 2011]. Dans ce contexte, un système automatisé de fouille de points de vue pourrait être employé pour permettre aux utilisateurs d'accéder à des opinions variées (par exemple, sous forme de résumés de points de vue) et ainsi réduire l'effet des chambres d'échos et des bulles de filtre. Une autre application possible est la détection de fausses nouvelles (*fake news*) – un phénomène qui a fait couler beaucoup d'encre depuis l'élection présidentielle américaine de 2016 en raison de l'impact sur la victoire de Donald Trump qui lui est parfois imputé [Allcott et Gentzkow, 2017]. Les fausses nouvelles sont caractérisées par une forte subjectivité, ce qui justifie le potentiel de techniques basées sur la fouille de points de vue [Jin *et al.*, 2016].

Ainsi, la fouille de points de vue constitue une problématique chargée de défis scientifiques et présente un potentiel d'application à différents problèmes d'actualité. Dans la Section 2,

nous définissons plus spécifiquement les problématiques liées à la fouille de point de vue auxquelles nous nous sommes intéressé dans cette thèse.

2 Problématiques

Cette thèse attaque un sous-problème de la fouille de points de vue, que nous dénommerons « découverte de points de vue », dont l'objectif est d'identifier le point de vue exprimé dans chaque texte ou par chaque auteur (par exemple un utilisateur de médias sociaux) d'une collection de documents. La découverte de points de vue constitue la première étape vers la mise en œuvre des différentes applications sus-mentionnées et est par conséquent fondamentale. Ce problème pourrait être envisagé comme un problème de classification (chaque classe correspondant à un point de vue) et abordé par des approches supervisées. Cependant, l'annotation manuelle de collections de données – nécessaire aux approches supervisées – est souvent difficile et coûteuse à obtenir. Nous avons donc choisi dans cette thèse d'adopter un cadre non supervisé, dans lequel nous ne disposons pas d'informations *a priori* sur la nature des points de vue à identifier. Contrairement aux techniques non supervisées de fouille d'opinions classique – focalisées sur les opinions positives et négatives – nous ne pouvons pas exploiter des lexiques de mots pré-existants. En effet, le vocabulaire propre à un point de vue n'est généralement pas transposable à d'autres sujets. Par exemple, les mots spécifiques aux partisans de Donald Trump ou à ceux de Hillary Clinton ne pourront vraisemblablement pas être exploités pour différencier des textes écrits selon les points de vue pro-israéliens ou pro-palestiniens.

Par conséquent, nous explorons dans cette thèse d'autres indices disponibles dans les textes et sur les médias sociaux pour faciliter la découverte de points de vue. Les problématiques auxquelles nous nous sommes intéressé peuvent ainsi être découpées suivant les deux questions et quatre sous-questions suivantes :

1. Comment exploiter le contenu textuel des documents sur le Web pour découvrir les points de vue qui y sont exprimés ?
   (a) Comment utiliser à cette fin la co-occurrence des mots et la nature distributionnelle du langage, pour faire émerger des motifs lexicaux révélateurs des différents points de vue ?
   (b) Est-il possible et pertinent d'intégrer des indicateurs de subjectivité, tels que les parties de discours pour identifier les mots porteurs de points de vue ?
2. Comment tirer parti des métadonnées disponibles dans les médias sociaux, telles que les interactions entre utilisateurs, pour raffiner la découverte de points de vue ?
   (a) Une approche exploitant conjointement le contenu textuel et les métadonnées est-elle envisageable ?
   (b) Quel est l'impact de ces métadonnées sur les performances d'une telle approche, en comparaison avec la seule exploitation d'indices textuels ?

Les contributions que nous avons apportées dans le cadre de cette thèse pour répondre à ces questions sont synthétisées dans la Section 3.

3 Contributions

Comme nous l'avons mentionné dans la Section 2, une caractéristique des documents textuels qui s'avère précieuse dans le cadre non supervisé est la co-occurrence des mots : deux mots qui apparaissent fréquemment dans le même contexte (par exemple, le même document ou la même phrase) ont tendance à être sémantiquement liés. Ce phénomène peut être exploité par une catégorie de modèles non supervisés nommés modèles thématiques (*topic models*), dans laquelle l'allocation de Dirichlet latente (LDA – *latent Dirichlet allocation*) [Blei *et al.*, 2001, 2003] s'inscrit. Les modèles thématiques constituent une approche polyvalente pour découvrir sans annotations préalables des thèmes ainsi que d'autres dimensions latentes (telles que les points de vue) à partir d'un corpus de textes. Nous avons donc décidé de baser nos approches pour la fouille de points de vue sur les modèles thématiques et en particulier ceux inspirés de LDA. Nos contributions à la fouille de points de vue reposent sur la proposition de deux modèles thématiques originaux :

1. Notre première contribution [Thonet *et al.*, 2016] se focalise sur la modélisation des points de vue dans les documents textuels sans métadonnées disponibles. Nous définissons en particulier la tâche de *découverte des points de vue et des opinions* (*viewpoint and opinion discovery*), qui consiste à analyser une collection de documents afin d'identifier le point de vue de chaque document, les thèmes mentionnés par chaque document et les opinions spécifiques aux points de vue pour chaque thème. Pour traiter ce problème, nous proposons le modèle VODUM (*viewpoint and opinion discovery unification model*), une approche non supervisée permettant la modélisation conjointe des points de vue et des thèmes en différenciant mots d'opinion (spécifiques aux points de vue et aux thèmes) et mots thématiques (spécifiques aux thèmes uniquement et neutres vis-à-vis des différents points de vue). À travers VODUM, nous étudions dans quelle mesure les *parties de discours* peuvent être exploitées pour distinguer les mots d'opinion et les mots thématiques dans un cadre non supervisé.

2. Dans notre seconde contribution [Thonet *et al.*, 2017], nous nous intéressons à étendre la découverte de points de vue aux réseaux sociaux en exploitant les métadonnées qui y sont disponibles. En particulier, notre objectif est ici d'analyser dans quelle mesure l'utilisation des interactions entre utilisateurs, en plus de leur contenu textuel généré, est bénéfique pour l'identification de leurs points de vue. L'intuition que nous développons ici repose sur le principe d'homophilie selon lequel les individus « similaires » entre eux (par exemple dans leurs opinions politiques) ont une plus forte propension à créer des liens. Nous présentons ainsi dans un premier temps le modèle SNVDM (*Social Network Viewpoint Discovery Model*) qui exploite conjointement le contenu généré par les utilisateurs et leurs interactions pour modéliser sans supervision à la fois les points de vue et les thèmes qui leur sont associés. Afin de surmonter les cas où le réseau d'interactions sociales est peu dense (*sparse*) – lorsqu'un utilisateur n'interagit qu'avec un nombre limité d'utilisateurs tiers – nous proposons ensuite une extension de SNVDM, nommée SNVDM-GPU, basée sur les urnes de Pólya généralisées [Mahmoud, 2008]. SNVDM-GPU présente notamment l'avantage d'intégrer les relations d'« accointances d'accointances » afin de prendre en compte les liens faibles existant entre les utilisateurs.

4 Organisation du mémoire

Le contenu de ce mémoire est organisé en deux parties. La Partie I synthétise les travaux de l'état de l'art pertinents aux problématiques abordées dans cette thèse. En particulier, le Chapitre 1 décrit les concepts de base, les tâches et les méthodes associées à la fouille d'opinions. Ce chapitre couvre l'approche « classique » de la fouille d'opinions, focalisée sur les opinions positives et négatives. Dans le Chapitre 2, nous introduisons le sous-domaine de la fouille d'opinions que nous dénotons par l'appellation « fouille de points de vue ». Nous proposons d'unifier sous cette dénomination un ensemble de tâches et de scénarios reliés par la notion de point de vue. Le Chapitre 3 détaille quant à lui l'allocation de Dirichlet latente (LDA – *latent Dirichlet allocation*), que nos contributions proposent d'étendre afin de modéliser conjointement thèmes et points de vue. Plus généralement, ce chapitre décrit les aspects méthodologiques et les éléments mathématiques inhérents aux modèles thématiques, nécessaires à la contribution des chapitres suivants.

La Partie II présente ensuite les contributions de cette thèse, œuvrant à faciliter la découverte de points de vue sur le Web lorsqu'aucune donnée annotée n'est disponible. Notre première contribution, détaillée dans le Chapitre 4, définit une modélisation conjointe des thèmes et des points de vue dans laquelle nous explorons l'idée de différencier mots d'opinions (spécifiques à la fois à un point de vue et à un thème) et mots thématiques (dépendants du thème mais neutres vis-à-vis des différents points de vue). Cette séparation entre mots d'opinion et mots thématiques est basée sur les parties de discours, inspirée par des pratiques similaires dans la littérature de fouille d'opinions classique – restreinte aux opinions positives et négatives. Dans le Chapitre 5, nous nous focalisons cette-fois sur les points de vue exprimés sur les réseaux sociaux. Notre objectif est alors d'analyser dans quelle mesure l'utilisation des interactions entre utilisateurs, en plus de leur contenu textuel généré, est bénéfique pour l'identification de leurs points de vue.

Enfin, nous concluons le mémoire en résumant les différents résultats obtenus dans cette thèse et en discutant des extensions possibles de nos travaux.

Première partie

Revue de l'état de l'art

Chapitre 1

Fouille d'opinions

Sommaire

1.1 Introduction

Nous présentons dans ce chapitre un aperçu des travaux existants en fouille d'opinions. Étant donné le volume important de littérature en la matière, cet état de l'art sur la fouille d'opinions n'a pas vocation à être exhaustif. Il en couvre les concepts et problèmes principaux afin de fournir le contexte de la fouille de points de vue – le cœur de cette thèse – qui constitue un sous-domaine de la fouille d'opinions. Pour une revue plus exhaustive et détaillée sur la fouille d'opinions, le lecteur intéressé peut se référer aux deux états de l'art références dans ce domaine : celui de Pang et Lee [2008] et celui de Liu [2012].

Tout d'abord, nous définirons les concepts clés et les objectifs de la fouille d'opinions (Section 1.1.1). Dans les Sections 1.1.2 et 1.1.3 nous expliquerons dans quelle mesure les systèmes de fouille d'opinions sont critiques autant pour les individus que pour les entreprises et les décideurs, et quels sont les défis de tels systèmes, respectivement. La Section 1.1.4 donnera un aperçu des tâches successives nécessaires à la construction d'un système de fouille d'opinions : la détection de la subjectivité, l'identification de la polarité des opinions, et la génération de résumés d'opinions, détaillées respectivement dans les Sections 1.2, 1.3 et 1.4.

1.1.1 Définitions et objectifs

La fouille d'opinions (*opinion mining*), également parfois désignée sous le nom d'analyse de sentiments (*sentiment analysis*), est un sous-domaine de l'informatique à l'intersection de plusieurs disciplines telles que le traitement automatique du langage naturel, la recherche d'information, la fouille de texte et l'apprentissage automatique. Les termes « fouille d'opinions » et « analyse de sentiments » ont été respectivement introduits dans [Dave *et al.*, 2003] et [Nasukawa et Yi, 2003]. Avant de définir plus précisément ce sous-domaine, attardons-nous d'abord sur la notion clé d'*opinion*. Selon le dictionnaire Larousse en ligne [1], une opinion désigne :

1. un « jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense » ;
2. ou l'« ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc. ».

La première définition positionne l'opinion au niveau de l'individu, alors que la seconde définition évoque une opinion collective et partagée. Considérons les exemples suivants pour apprécier les nuances de ces deux définitions :

1. Quelle est ton opinion sur ce livre ?
2. Quelles sont tes opinions politiques ?

La première question attend principalement une réponse exprimant un avis positif ou négatif (voire neutre éventuellement) exprimant respectivement si la personne questionnée a aimé ou non le livre évoqué. Il s'agit du type d'opinions personnelles ciblées (ici, vers un livre) que l'on peut trouver dans les critiques de produits ou de services en ligne (*online reviews*) sur des sites web tels que Amazon [2] et TripAdvisor [3]. La seconde question est quant à elle plus complexe et demande une réponse élaborée, autre que positive ou négative. Il est attendu de la personne questionnée qu'elle se positionne par rapport au monde et à la société en définissant un ensemble de principes et de valeurs – qui sont par ailleurs communs aux individus partageant son idéologie politique. La première question fait ainsi référence à la première définition de l'opinion – celle d'une opinion individuelle – alors que la seconde question mentionne la seconde définition – celle d'une opinion collective. Cette seconde définition correspond en réalité à ce que nous nommerons plus tard « points de vue » dans le Chapitre 2.

La notion d'opinion étant clarifiée, nous pouvons maintenant expliciter la tâche de fouille d'opinions. Dans leur état de l'art qui fait référence en la matière [Pang et Lee, 2008], Pang et Lee définissent la fouille d'opinions comme « le traitement informatique de l'opinion, du sentiment et de la subjectivité dans le texte ». Ici, le « traitement informatique » se rapporte à un processus automatisé et algorithmique. « Sentiment » peut être considéré comme un synonyme de l'opinion individuelle détaillée dans la première définition ; ce terme ne doit pas être confondu avec ses définitions alternatives telles que l'état affectif ou le penchant [4]. Un

1. http://www.larousse.fr/dictionnaires/francais/opinion/56197
2. https://www.amazon.fr/
3. https://www.tripadvisor.fr/
4. http://www.larousse.fr/dictionnaires/francais/sentiment/72138

texte est considéré comme subjectif lorsqu'il exprime une opinion – nous reviendrons sur la notion de subjectivité dans la Section 1.2.

Les travaux en fouille d'opinions se sont longtemps focalisés sur les opinions individuelles formulées dans les critiques en ligne. Ce n'est que plus récemment, depuis des travaux pionniers tels que [Lin *et al.*, 2006, 2008; Paul *et al.*, 2010; Popescu et Pennacchiotti, 2010], que les chercheurs ont étudié les opinions collectives (par exemple, les opinions politiques), désormais exprimées massivement dans les médias sociaux. Dans le présent chapitre, nous nous intéresserons essentiellement à ces premiers travaux sur les opinions individuelles, qui ont établi les bases de la fouille d'opinions. Les travaux sur l'analyse de points de vue (ou, autrement dit, l'analyse d'opinions collectives) – qui sont au cœur du problème étudié dans cette thèse – seront abordés en détail dans le Chapitre 2. Par conséquent, sauf en cas de mention explicite du contraire, nous utiliserons simplement « fouille d'opinions » pour désigner la fouille d'opinions individuelles dans le reste de ce chapitre.

1.1.2 Motivations et applications

Depuis l'apparition du phénomène populairement nommé « Web 2.0 » [5], les internautes se sont vus offrir de nouvelles possibilités en matière d'interaction et de sociabilité en ligne. Les nouveaux services proposés permettent aux utilisateurs d'Internet de générer leur propre contenu et ainsi exprimer leurs opinions, par exemple sous la forme de billets de blog, ou encore par l'intermédiaire de posts sur des plateformes de réseaux sociaux telles que Twitter [6] et Facebook [7]. Ainsi, en 2016, le nombre d'utilisateurs de médias sociaux était estimé à 2,34 milliards dans le monde [8]. D'après un rapport du Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie (CRÉDOC [9]) de 2016 [Croutte et Lautié, 2016], le pourcentage d'internautes français membres de réseaux sociaux s'élevait à 56 % sur l'ensemble des classes d'âge, et 84 % pour les moins de 40 ans. Les critiques publiées sur les sites de commerce en ligne tels que Amazon et TripAdvisor jouent également un rôle prépondérant pour les consommateurs. Selon un sondage du *Pew Research Center* [10] réalisé en 2016 auprès de 4 787 adultes américains, 82 % des personnes interrogées ont déclaré consulter les notes et les critiques en ligne lorsqu'elles achètent un produit pour la première fois [Smith et Anderson, 2016]. De plus, 39 % ont déjà partagé leur opinion vis-à-vis d'un produit dans les médias sociaux.

Analysons un exemple pour montrer l'intérêt des systèmes de fouille d'opinions vis-à-vis des critiques en ligne. La Figure 1.1 montre un extrait des critiques de la tablette Fire rédigées sur Amazon. [11] Un total de 859 critiques en français ont été postées par les utilisateurs d'Amazon qui ont acheté ce produit. On peut observer qu'Amazon indique par un histogramme les

5. http://www.oreilly.com/pub/a//web2/archive/what-is-web-20.html
6. https://twitter.com/
7. https://www.facebook.com/
8. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
9. http://www.credoc.fr/
10. http://www.pewresearch.org/
11. https://www.amazon.fr/Amazon-Tablette-Fire-7-Pouces-8Go/dp/B00ZDWLEEG/

FIGURE 1.1 – Exemple de critiques postées sur Amazon (11 juin 2016).

proportions des différentes notes attribuées (entre une et cinq étoiles). On note également qu'une critique positive et une critique négative sont mises en avant en haut de la page. Le reste de la page contient ensuite la liste de toutes les critiques, qu'il est éventuellement possible de trier par note. L'acheteur potentiel qui ne se contente pas de la note moyenne et souhaite prendre en compte les critiques pour prendre sa décision d'achat se verra contraint de lire une par une un grand nombre de ces critiques afin d'avoir une idée précise de la qualité du produit. Cette lecture s'avère rébarbative et coûteuse en temps.

Par conséquent, un système capable de fournir automatiquement un résumé des différentes opinions exprimées sur un produit permettrait de réduire considérablement les efforts de l'utilisateur. Un tel résumé consisterait par exemple à présenter les aspects positifs et négatifs du produit ciblé par les critiques – nous reviendrons sur la notion d'aspect dans la Section 1.3.2. Un tel résumé n'est pas seulement utile pour les utilisateurs, il l'est aussi pour l'entreprise qui a fabriqué le produit : savoir quels aspects du produit ont été appréciés ou non permet de corriger ses défauts ou en proposer une version améliorée dans le futur. Plus généralement, le contenu généré par les internautes vis-à-vis d'un produit – dans des critiques en ligne, sur des blogs ou sur les réseaux sociaux – permet à l'entreprise concernée de surveiller l'image publique de sa marque (faire du *brand monitoring*) [Kim, 2006]. Gérer sa réputation est également un élément clé pour les acteurs politiques.

Ainsi, le développement de systèmes de fouille d'opinions est une tâche critique avec un impact à la fois économique et politique. Cette raison a motivé l'intérêt des chercheurs – en particulier dans le domaine du traitement automatique du langage naturel et celui de la recherche d'information – et la nécessité d'évaluer ces systèmes a mené au développement de *benchmarks* sur la fouille d'opinions dans des campagnes d'évaluation telles que TREC [12] (*Text REtrieval Conference*) en 2006 [Ounis *et al.*, 2006], et SemEval [13] (*Semantic Evaluation Workshop*) entre 2007 et 2017 [Kiritchenko et Mohammad, 2016; Mohammad *et al.*, 2016; Nakov *et al.*, 2016, 2013; Pontiki *et al.*, 2016, 2014; Rosenthal *et al.*, 2015; Strapparava et Mihalcea, 2007; Wu, 2010]. Dans la section qui suit, nous décrivons les défis que présente la construction d'un système de fouille d'opinions.

1.1.3 Défis

Étant donné que la fouille d'opinions est une instance de la fouille de textes, il est légitime de se demander ce qui rend cette première tâche spécifique et difficile. Par exemple, qu'est-ce qui différencie la classification de textes positifs et négatifs de la classification de courriels indésirables (*spam*) et de courriels désirables (*ham*) [Russell et Norvig, 2010] ? Le niveau de difficulté d'une tâche de classification de textes est lié aux différences lexicales entre les classes : plus les classes utilisent un vocabulaire distinct, plus il sera facile d'assigner un texte à la bonne classe. Les courriels indésirables sont souvent des publicités, par exemple pour des sites pornographiques ou des médicaments tels que le viagra. Le lexique de ce genre de contenu est généralement distinct de celui des courriels désirables, traitant par exemple de

12. `http://trec.nist.gov/`
13. `https://en.wikipedia.org/wiki/SemEval`

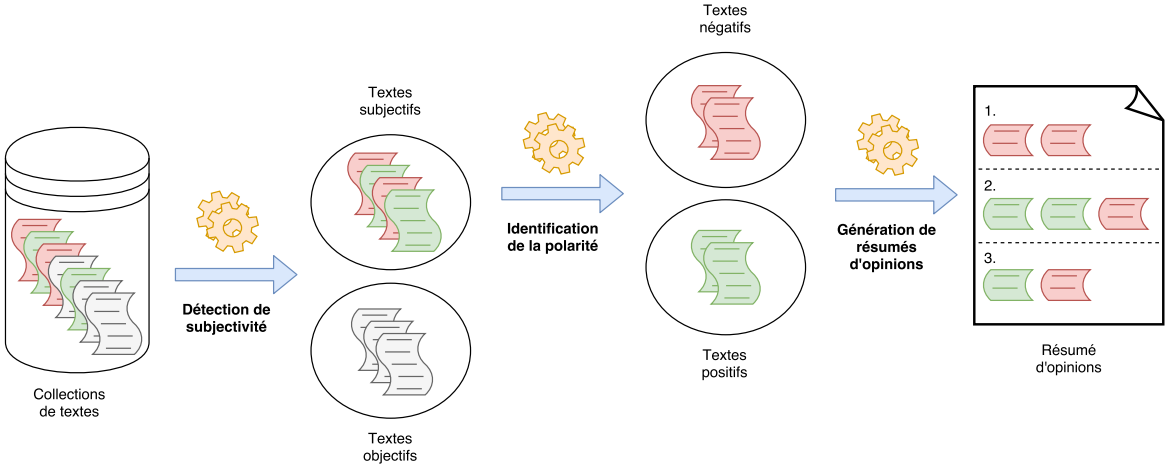


FIGURE 1.2 – Processus complet de fouille d'opinions.

sujets professionnels. Cependant, la différence entre un texte d'opinion positive et un texte d'opinion négative est plus subtile. Il est possible que les deux textes traitent du même thème (par exemple, un film de cinéma donné), induisant ainsi une grande similarité lexicale entre les classes positives et négatives. De plus, bien que certains mots tels que « super » et « mauvais » semblent indiquer de manière fiable la classe d'opinions, ignorer la négation de ces adjectifs faussera la classification. Par ailleurs, une opinion peut être exprimée de manière implicite (« Je ne reviendrai pas dans ce restaurant. ») ou peut même inclure de l'ironie (« Bravo les bleus pour votre excellent match ! ») [Karoui *et al.*, 2015]. Ainsi, la fouille d'opinions nécessite d'être abordée différemment de la fouille de textes classiques.

1.1.4 Processus de fouille d'opinions

La fouille d'opinions peut être considérée comme un processus en plusieurs étapes qui prend en entrée un ensemble de textes sur une cible (*target* – par exemple, un produit ou une personnalité) et fournit en sortie un résumé agrégeant les opinions exprimées dans le texte vis-à-vis de la cible ou éventuellement vis-à-vis des aspects de la cible [Dave *et al.*, 2003]. L'aspect d'une cible est une caractéristique, un attribut ou un élément composant de la cible ; nous reviendrons plus en détail sur cette notion dans la Section 1.3. On peut ainsi découper le processus de fouille d'opinions en trois étapes principales, illustrées dans la Figure 1.2 :

1. détecter les textes subjectifs, c'est-à-dire les textes exprimant une opinion vis-à-vis d'une cible donnée ;
2. identifier si l'opinion exprimée dans chaque texte subjectif est positive ou négative à l'égard de la cible et éventuellement à l'égard de ses aspects ;
3. former un résumé des différentes opinions présentes dans l'ensemble de textes.

Notons que dans ce processus nous supposons avoir déjà à disposition un ensemble de textes pertinents vis-à-vis de la cible. Obtenir un tel ensemble de textes, par exemple à partir d'une requête formulée en langage naturel, est généralement loin d'être trivial – c'est le cœur même

de la tâche de la recherche d'information. Cependant, ce type de pré-filtrage basé sur la pertinence ne nécessite pas de traitement propre aux textes d'opinions. Nous n'aborderons donc pas cette phase de recherche des textes pertinents dans ce mémoire. Pour un aperçu plus détaillé sur les techniques utilisées en recherche d'information, le lecteur peut se référer aux ouvrages de référence de Baeza-Yates et Ribeiro-Neto [1999] et de Manning *et al.* [2008].

Dans le reste de ce chapitre, nous détaillons les travaux de l'état de l'art associés à chacune des trois étapes du processus de fouille d'opinions. La Section 1.2 présente les approches permettant de détecter si un texte est subjectif. Dans la Section 1.3, nous décrivons les différentes méthodes d'identification de la polarité des opinions (c'est-à-dire l'orientation positive ou négative des opinions). Enfin, la Section 1.4 passe en revue les travaux sur la génération de résumés d'opinions à partir des opinions extraites à l'issue des étapes précédentes.

1.2 Détection de subjectivité

Une tâche préliminaire à l'analyse des opinions contenues dans une collection de textes consiste à détecter les documents ou portions de documents subjectifs, c'est-à-dire exprimant des opinions. En effet, certains documents peuvent s'avérer purement factuels (par exemple, un article de presse relatant un évènement sportif) alors que d'autres documents mentionnant des sujets plus polémiques reflètent les opinions de leurs auteurs (par exemple, un essai politique). De plus, un document exprimant une opinion n'est pas nécessairement subjectif dans son intégralité. Par exemple, une critique en ligne sur un téléphone mobile peut contenir une phrase telle que « J'ai commandé le modèle blanc. », qui ne porte aucune marque d'opinion et pourrait aussi bien être utilisée dans une critique positive que dans une critique négative.

La détection de subjectivité constitue ainsi un problème à part entière – qui se révèle en réalité être souvent plus difficile que l'analyse subséquente de la polarité des opinions [Mihalcea *et al.*, 2007]. Afin d'encourager la recherche sur ce problème, la campagne d'évaluation TREC a proposé en 2006 une tâche de recherche d'opinions sur les blogs [Ounis *et al.*, 2006]. Dans le cadre de cette tâche, un document est jugé subjectif s'il contient « une expression explicite d'opinion ou de sentiment vis-à-vis de la cible, révélant une position personnelle de l'auteur » (traduit de l'anglais). Le but de TREC Blog 2006 était ainsi d'identifier les documents (c'est-à-dire les posts de blogs) à la fois pertinents vis-à-vis d'un sujet donné et subjectifs. D'autres travaux ont porté sur la détection de subjectivité au niveau de la phrase [Hatzivassiloglou et Wiebe, 2000; Wiebe *et al.*, 1999; Yu et Hatzivassiloglou, 2003] ou au niveau des expressions [Riloff et Wiebe, 2003; Wiebe et Wilson, 2002] plutôt qu'au niveau du document. De manière générale, les approches pour la détection de subjectivité reposent sur une combinaison des méthodes suivantes :

— l'utilisation de lexiques de mots d'opinions externes, construits manuellement ou automatiquement [Mishne, 2006; Oard *et al.*, 2006; Yang *et al.*, 2006] ;

— l'exploitation de marqueurs linguistiques tels que les parties du discours (*part of speech*) en considérant par exemple les pronoms et les adjectifs comme des marques de subjec-

tivité [Hatzivassiloglou et Wiebe, 2000; Riloff et Wiebe, 2003; Wiebe et Wilson, 2002; Yang *et al.*, 2006; Yu et Hatzivassiloglou, 2003] ;

— la mise en œuvre de classifieurs supervisés tels que les machines à vecteurs de support (SVM) et classifieurs bayésiens naïfs [Hatzivassiloglou et Wiebe, 2000; Wiebe et Wilson, 2002; Wiebe *et al.*, 1999; Yu et Hatzivassiloglou, 2003; Zhang et Yu, 2006] ;

— l'application de méthodes symboliques basées sur des règles et des motifs, définis manuellement ou automatiquement [Riloff et Wiebe, 2003; Wiebe et Wilson, 2002].

Une fois que les documents ou fragments de texte subjectifs ont été détectés, les opinions qui y sont exprimées peuvent en être extraites et leur polarité identifiée.

1.3 Identification de la polarité et de la nuance des opinions

En fouille d'opinions individuelles, les opinions sont considérées comme positives, négatives ou une nuance de ces extrêmes. On désigne ainsi par *degré de polarité* ou simplement *polarité* la position d'une opinion sur cet axe comprenant les différents niveaux de négativité et de positivité – on trouve également le terme « orientation sémantique » dans la littérature [Hatzivassiloglou et McKeown, 1997]. La polarité peut être définie par des catégories telles que « positif », « neutre », et « négatif », ou encore par des nombres dénotant le degré de positivité ou de négativité. Par exemple, la polarité peut être définie entre 1 et 5, où 1 désigne une polarité très négative, et 5 désigne une polarité très positive – cela correspond au format des notes données dans les critiques en ligne sur Amazon et TripAdvisor.

Comme nous l'avons précisé dans la Section 1.1.1, une opinion individuelle est ciblée. Cette cible peut être de diverses natures suivant le type de données d'opinions concernées. Par exemple, une critique en ligne cible généralement un produit commercial (par exemple, un téléphone mobile) ou un service (par exemple, un hébergement dans un hôtel). Un message sur un réseau social ou sur un blog peut quant à lui porter sur une célébrité, telle qu'un artiste ou un homme politique. Ainsi, on peut considérer qu'un texte d'opinion dont la polarité est positive révèle l'approbation globale de l'auteur du texte vis-à-vis de la cible, et inversement un texte négatif dénote une dépréciation globale. Cette considération présuppose implicitement qu'un texte d'opinion est homogène : il est soit totalement positif, soit totalement négatif (éventuellement neutre) vis-à-vis de la cible étudiée. La Section 1.3.1 détaille les travaux basés sur ce postulat en considérant les opinions à un niveau de granularité grossière (*coarse-grained opinions)*, c'est-à-dire en étudiant les opinions associées à la cible dans sa globalité.

1.3.1 Cas des opinions de granularité grossière

Les premières approches de fouille d'opinions s'intéressent à l'opinion globale exprimée vis-à-vis de la cible dans un document subjectif. Autrement dit, ces travaux considèrent qu'un document est associé à une polarité unique et qu'un document ne forme pas un mélange d'opinions de polarités différentes. Cette considération est d'autant plus valide que le document

est court. Par exemple, une critique en ligne témoigne généralement de l'appréciation globale ou de la dépréciation globale du produit ou service.

À partir de cette supposition, de nombreux travaux se sont donnés pour but l'identification automatique de la polarité globale au niveau du document [Dave *et al.*, 2003; Gamon, 2004; Goldberg et Zhu, 2006; McDonald *et al.*, 2007; Pang et Lee, 2004, 2005; Pang *et al.*, 2002; Turney, 2002] ou de la phrase [Kim et Hovy, 2004; McDonald *et al.*, 2007; Socher *et al.*, 2011; Yu et Hatzivassiloglou, 2003]. Ces différentes approches peuvent être regroupées en fonction du niveau de supervision adopté et la nature du problème d'apprentissage de l'identification de polarité :

— classification supervisée : [Dave *et al.*, 2003; Gamon, 2004; Kim et Hovy, 2004; McDonald *et al.*, 2007; Pang et Lee, 2004; Pang *et al.*, 2002; Yu et Hatzivassiloglou, 2003] ;

— classification semi-supervisée ou non supervisée : [Socher *et al.*, 2011; Turney, 2002] ;

— régression : [Goldberg et Zhu, 2006; Pang et Lee, 2005].

Nous détaillons dans le reste de cette section les travaux adoptant ces différents paradigmes.

1.3.1.1 Classification supervisée

Les méthodes de classification supervisée abordent le problème d'identification de polarité en se basant sur deux catégories (positif et négatif) [Dave *et al.*, 2003; Gamon, 2004; Pang et Lee, 2004; Pang *et al.*, 2002; Yu et Hatzivassiloglou, 2003] ou trois catégories (positif, négatif, ou neutre) [Kim et Hovy, 2004]. Elles sont basées sur des classifieurs tels que le classifieur bayésien naïf [Dave *et al.*, 2003; Pang et Lee, 2004; Pang *et al.*, 2002], la machine à vecteurs de support [Gamon, 2004; Pang et Lee, 2004; Pang *et al.*, 2002] ou le classifieur à maximum d'entropie [Pang *et al.*, 2002]. McDonald *et al.* [2007] proposent quant à eux de traiter le problème de classification en utilisant un modèle apparenté aux champs aléatoires conditionnels (*conditional random fields*). Les traits (*features*) des différents classifieurs sont essentiellement basés sur les n-grammes [Dave *et al.*, 2003; Gamon, 2004; McDonald *et al.*, 2007; Pang *et al.*, 2002], les parties du discours [Gamon, 2004; McDonald *et al.*, 2007; Pang *et al.*, 2002], la position relative des mots [Dave *et al.*, 2003; Pang *et al.*, 2002], ou sur des ressources externes telles que WordNet [Dave *et al.*, 2003] ou des lexiques de mots d'opinions [Yu et Hatzivassiloglou, 2003]. Dans [Pang et Lee, 2004], l'identification de polarité est réalisée à partir de traits extraits des phrases subjectives uniquement, obtenues par formulation d'un problème de graphes basé sur la coupe minimale (*minimum cut*).

1.3.1.2 Classification semi-supervisée ou non supervisée

L'inconvénient majeur des classifieurs supervisés est leur dépendance vis-à-vis d'une quantité importante d'exemples annotés, utilisés lors de la phase d'apprentissage. Ces données annotées sont généralement difficiles à obtenir et peuvent représenter un coup prohibitif lorsque l'annotation doit être réalisée manuellement. Par conséquent, l'identification de la

polarité d'un texte a également été abordé sous l'angle d'un problème de classification semi-supervisée [Socher *et al.*, 2011], voire non supervisée [Turney, 2002]. L'approche semi-supervisée proposée dans [Socher *et al.*, 2011] est basée sur un réseau de neurones auto-encodeur récursif (*recursive autoencoder*) pour prédire la distribution d'opinions au niveau de la phrase. Elle tire avantage de la nature compositionnelle de la sémantique en déduisant la polarité d'une phrase à partir de celle des mots qui la composent. L'algorithme décrit par Turney [2002] est quant à lui totalement non supervisé. Il estime dans un premier temps la polarité des adjectifs et des adverbes présents dans des critiques d'opinion en calculant leur proximité, basée sur l'information mutuelle ponctuelle (PMI – *pointwise mutual information*), avec des mots tels que *poor* et *excellent*. La polarité globale de la critique est ensuite déduite en agrégeant la polarité des adjectifs et adverbes qui la composent.

1.3.1.3 Régression

Une vision alternative de la tâche d'identification de la polarité est de la considérer comme un problème de régression. Désormais, le but n'est plus d'assigner au texte d'opinion une catégorie parmi {positif, négatif} ou parmi {positif, négatif, neutre}, mais plutôt de lui associer un nombre (éventuellement réel) dénotant le degré de polarité du texte. En adoptant le système de notes utilisé sur Amazon et TripAdvisor (entre 1 et 5), l'objectif des modèles de régression est de prédire les notes des différentes critiques. Une telle approche a été proposée par Pang et Lee [2005], basée sur une régression par machine à vecteurs de support. La fonction de similarité entre deux textes, nécessaire à la tâche de régression, est définie à partir de la proportion de phrases positives et de phrases négatives dans les textes, apprise par un classifieur bayésien naïf entraîné sur un corpus externe de critiques positives et négatives. Ce travail a par la suite été étendu dans [Goldberg et Zhu, 2006] où est présentée une approche semi-supervisée basée sur une représentation des documents sous forme de graphe, permettant ainsi d'exploiter des données d'apprentissage non annotées.

Les approches présentées dans cette section suppose qu'un texte d'opinion, tel qu'un document ou une phrase, reflète une opinion de polarité unique vis-à-vis de la cible étudiée. En réalité, un tel texte est souvent nuancé et n'exprime pas uniquement une opinion globalement positive ou globalement négative. Par exemple, le spectateur d'un film de cinéma peut expliquer dans une critique qu'il trouve l'histoire intéressante et le jeu d'acteurs de qualité, mais que les décors sont de mauvaise facture et les effets spéciaux pauvres. On voit émerger à travers cet exemple la notion d'*aspect* : un aspect d'un texte d'opinion est une caractéristique ou un attribut de la cible sur lequel l'auteur a émis un avis. Dans l'exemple précédent, la cible est le film de cinéma et les aspects sont l'histoire, le jeu d'acteurs, les décors et les effets spéciaux. Cette approche de l'opinion à un niveau de granularité plus fin est connue dans la littérature sous le nom de fouille d'opinions basées sur les aspects (*aspect-based opinion mining*). La Section 1.3.2 décrit les différents travaux identifiant la polarité des opinions basées sur les aspects.

1.3.2 Cas des opinions basées sur les aspects

Pour une cible donnée, l'identification de la polarité des opinions basées sur des aspects consiste à déterminer la polarité associée à chaque aspect de la cible. Cependant, les aspects d'une cible ne sont généralement pas connus *a priori*, et ils varient d'une cible à une autre. Par exemple, les aspects d'un téléphone mobile sont sa batterie, son appareil photo, son poids, etc., alors que les aspects d'un film de cinéma sont son histoire, son jeu d'acteurs, ses décors, etc. Ainsi, une difficulté additionnelle dans la fouille d'opinions basées sur des aspects est d'extraire les aspects dans un premier temps. La seconde étape, similaire à l'identification de la polarité d'opinions de granularité grossière, associe une polarité aux différents aspects extraits.

Les opinions basées sur des aspects ont été le sujet d'un important nombre de travaux et ont également fait l'objet de plusieurs tâches dans la campagne d'évaluation SemEval [Pontiki *et al.*, 2016, 2014]. On distingue dans la littérature deux types de travaux sur la fouille d'opinions basées sur les aspects :

— Certains travaux séparent le problème en deux phases et traitent l'une de ces phases ou les deux : extraction des aspects de la cible dans le texte [Brody et Elhadad, 2010; Hu et Liu, 2004; Liu *et al.*, 2005; Popescu et Etzioni, 2005] et identification des opinions (ainsi que leur polarité) associées aux aspects [Brody et Elhadad, 2010; Ding *et al.*, 2008; Hu et Liu, 2004; Popescu et Etzioni, 2005; Snyder et Barzilay, 2007].

— D'autres travaux proposent de découvrir conjointement les aspects et les opinions [He *et al.*, 2012, 2013; Jo et Oh, 2011; Lim et Buntine, 2014; Lin et He, 2009; Lin *et al.*, 2012; Rahman et Wang, 2016; Wang *et al.*, 2016; Zhao *et al.*, 2010].

Nous détaillons dans le reste de cette section les approches permettant l'extraction des aspects, les approches identifiant les opinions exprimées vis-à-vis des aspects, et les approches proposant une solution unifiée à ces deux problèmes.

1.3.2.1 Extraction des aspects

L'extraction d'aspects peut être considérée comme une instance du problème d'extraction d'informations (*information extraction*) : l'objectif est d'inférer des informations structurées (la liste des aspects) à partir de données non structurées (les textes d'opinions). Certains travaux ont ainsi opté pour une approche symbolique basée sur des règles et sur les parties de discours [Hu et Liu, 2004; Liu *et al.*, 2005]. La méthode utilisée dans [Hu et Liu, 2004; Liu *et al.*, 2005] se base sur un algorithme de fouille d'associations (*association mining*) en s'appuyant sur l'hypothèse que les aspects sont souvent représentés par des syntagmes nominaux. De manière similaire, l'approche de Popescu et Etzioni [2005] extrait les aspects en calculant l'information mutuelle ponctuelle entre les syntagmes nominaux et la cible. Une méthode basée sur l'allocation de Dirichlet latente (*latent dirichlet allocation*) a également été proposée pour découvrir les aspects, en supposant l'équivalence entre thèmes et aspects [Brody et Elhadad, 2010].

1.3.2.2 Identification des opinions associées aux aspects

De même que pour l'extraction d'aspects, l'identification des opinions peut exploiter les parties de discours. De nombreux travaux associent l'expression de l'opinion aux adjectifs et aux adverbes [Brody et Elhadad, 2010; Ding *et al.*, 2008; Hu et Liu, 2004]. Ainsi, les mots d'opinions associés à un aspect peuvent être détectés en considérant les adjectifs et adverbes situés à proximité des mots dénotant des aspects (extraits lors de la phase précédente). Popescu et Etzioni [2005] adoptent une approche alternative basée sur un analyseur de dépendance (*dependency parser*) et un ensemble de règles définies manuellement.

Afin d'identifier la polarité des opinions extraites, Hu et Liu [2004] définissent tout d'abord manuellement une liste de 30 graines (*seed words*) clairement positifs ou négatifs (par exemple, *fantastic*, *cool*, *dull*, *bad*). Ensuite, la polarité des mots d'opinions est estimée en exploitant les relations de synonymie et d'antonymie issues de WordNet avec les germes. Une méthode similaire est mise en œuvre dans [Brody et Elhadad, 2010], où la polarité des mots d'opinions est identifiée à partir d'une liste de germes et l'application de l'algorithme de propagation d'étiquettes (*label propagation*). Ding *et al.* [2008] étendent les mots d'opinions aux noms et aux verbes en utilisant les germes proposés dans [Hu et Liu, 2004]. La négation est également prise en compte dans [Brody et Elhadad, 2010; Ding *et al.*, 2008] : la polarité d'une opinion est inversée à proximité d'un mot de négation. Dans [Popescu et Etzioni, 2005], l'identification de la polarité est effectuée en appliquant un algorithme de relaxation d'étiquettes (*label relaxation*) initialisé à partir de mots d'opinion classés par la méthode de Turney [2002].

Snyder et Barzilay [2007] ont quant à eux proposé une approche supervisée basée sur PRanking, un algorithme de perceptron en ligne. L'avantage de cette méthode est qu'elle ne nécessite pas l'identification préalable des mots d'opinions et elle est capable de prendre en compte la négation sans la définition manuelle de règles.

1.3.2.3 Approches conjointes

Plutôt que d'identifier les aspects et les opinions en deux étapes, de multiples travaux ont proposé de modéliser ces deux dimensions conjointement [He *et al.*, 2012, 2013; Jo et Oh, 2011; Kim *et al.*, 2013; Lim et Buntine, 2014; Lin et He, 2009; Lin *et al.*, 2012; Mei *et al.*, 2007; Moghaddam et Ester, 2011; Rahman et Wang, 2016; Wang *et al.*, 2016; Zhao *et al.*, 2010]. Pour ce faire, ces travaux se basent sur des approches de type modèle thématique probabiliste[14] (*probabilistic topic models*) tels que l'analyse sémantique latente probabiliste (*probabilistic latent semantic analysis* – PLSA) [Hofmann, 1999, 2001] ou l'allocation de Dirichlet latente (*latent Dirichlet allocation* – LDA) [Blei *et al.*, 2001, 2003]. Les approches basées sur les modèles thématiques supposent l'équivalence entre thème et aspect, et considèrent la polarité de l'opinion comme une dimension supplémentaire au thème. Les mots de polarité positive et les mots de polarité négative sont distingués en se basant sur des lexiques de

14. Nous donnerons une description plus générale des modèles thématiques probabilistes dans le Chapitre 3.

mots d'opinions tels que MPQA (*multi-perspective question answering*) [Wilson *et al.*, 2005] et SentiWordNet [Baccianella *et al.*, 2010].

Mei *et al.* [2007] s'inspirent de PLSA en utilisant un mélange de lois multinomiales afin de capturer à la fois la composante thématique et la composante d'opinion. Les travaux décrits dans [He *et al.*, 2012, 2013; Jo et Oh, 2011; Kim *et al.*, 2013; Lim et Buntine, 2014; Lin et He, 2009; Lin *et al.*, 2012; Moghaddam et Ester, 2011; Rahman et Wang, 2016; Wang *et al.*, 2016; Zhao *et al.*, 2010] étendent quant à eux le modèle LDA. Dans leur modèle précurseur JST (*joint sentiment/topic*), Lin et He [2009]; Lin *et al.* [2012] modifient LDA en ajoutant simplement une variable latente dénotant la polarité au niveau du mot – en plus de la variable latente dénotant le thème. Cette approche est par la suite adaptée dans [He *et al.*, 2012, 2013] pour intégrer la dimension temporelle et ainsi modéliser la dynamique à la fois thématique et d'opinions. Dans [Moghaddam et Ester, 2011], les mots d'opinions et les mots thématiques sont tout d'abord extraits par la méthode non supervisée décrite dans [Moghaddam, 2010], puis à partir de ces observations, le modèle identifie les notes (*ratings*) latentes – indicateurs similaires à la polarité dans le cadre des critiques en ligne – et les aspects latents. Lim et Buntine [2014] ont proposé un modèle inspiré de [Moghaddam et Ester, 2011] basé sur le processus de Pitman-Yor (*Pitman-Yor process*) et adapté à Twitter en intégrant les *hashtags*. Le modèle de Wang *et al.* [2016] fait quant à lui la distinction entre les mots dénotant un aspect, les mots traduisant une opinion générale et les mots exprimant une opinion spécifique à un aspect. De plus, afin d'intégrer la corrélation entre mots d'opinions issus de plusieurs domaines, les auteurs exploitent le modèle d'urnes de Pólya généralisées, qui étend le phénomène de co-occurrence des modèles thématiques.

Les modèles décrits précédemment ont choisi de modéliser les aspects (c'est-à-dire les thèmes) au niveau du mot. Une alternative est de supposer que chaque phrase se voit assigner un aspect unique. Il a en effet été observé qu'une phrase de critique en ligne, par exemple, contient souvent un unique aspect [Jo et Oh, 2011]. Cette hypothèse a été adoptée dans plusieurs modélisations conjointes des thèmes et des opinions [Jo et Oh, 2011; Kim *et al.*, 2013; Rahman et Wang, 2016; Zhao *et al.*, 2010]. Zhao *et al.* [2010] ont par ailleurs ajouté une dimension supervisée à leur approche par un modèle de maximum d'entropie afin de faciliter l'identification des aspects et des opinions. De manière similaire au modèle de Wang *et al.* [2016], le modèle proposé dans [Zhao *et al.*, 2010] distingue les mots vides de sens (*background words*), les mots d'opinions et d'aspects généraux, et les mots d'opinions et d'aspects spécifiques. L'approche de Rahman et Wang [2016] s'inspire quant à elle des modèles de Markov cachés en supposant une dépendance markovienne entre les thèmes des phrases successives d'une critique en ligne. L'intuition derrière cette dépendance est que le même aspect peut être discuté sur deux phrases successives, soit en conservant la même opinion, soit en nuançant l'opinion (c'est-à-dire en inversant la polarité). Plutôt qu'organiser tous les aspects à un même niveau de granularité, l'approche décrite dans [Kim *et al.*, 2013] forme automatiquement une hiérarchie d'aspects – et de sous-aspects, etc. – en adoptant une méthode similaire aux processus stochastiques des restaurants chinois emboîtés (*nested chinese restaurant process*).

Après l'identification des opinions d'un texte et de leur polarité (éventuellement vis-à-vis d'aspects), l'étape finale d'un système de fouille d'opinions consiste à organiser les opinions

sous forme de résumés (textuel ou non). Nous décrivons les travaux abordant cette tâche dans la Section 1.4.

1.4 Génération de résumés d'opinions

Dans un premier temps, définissons la tâche de génération de résumés textuels dans le cas général. Issue du domaine du traitement du langage naturel, la tâche de génération de résumés consiste à fournir, à partir d'un document (résumé individuel de document – *single-document summarization*) ou d'un ensemble de documents (résumé multi-documents – *multi-document summarization*), un texte court soulignant les informations centrales du ou des document(s). Cette tâche peut être abordée de deux manières différentes : en extrayant des phrases existantes dans le ou les document(s) (résumé extractif – *extractive summarization*) ou bien en formant de nouvelles phrases à partir des mots ou groupes de mots importants (résumé abstractif – *abstractive summarization*). Pour une description plus détaillée des travaux portant sur la génération de résumés dans le cas général, le lecteur peut se référer à l'un des multiples états de l'art existants, tels que [Nenkova et McKeown, 2011; Yao *et al.*, 2017].

La génération de résumés d'opinions est une instance de ce problème. Les résumés d'opinions se différencient des résumés généraux dans la mesure où ils doivent couvrir les éléments clés pour les opinions positives et négatives. Étant donné que la fouille d'opinions est fréquemment confrontée à un grand volume textes courts (tels que les critiques en ligne), la forme la plus courante de résumés d'opinions est le résumé multi-documents. Par exemple, à partir de toutes les critiques postées sur un film de cinéma, le but est d'en présenter succinctement et sans redondance les opinions positives et négatives pour le spectateur potentiel, qui cherche à décider s'il regardera ou non le film. Dans le reste de cette section, nous présentons les principaux travaux sur la génération de résumé d'opinions. Pour une plus grande exhaustivité sur le sujet, nous suggérons la lecture de la revue d'état de l'art rédigée par Kim *et al.* [2011].

1.4.1 Résumés d'opinions basés sur les aspects

Une première méthode pour la génération de résumés d'opinions consiste à prendre en compte la notion d'aspect, que nous avons introduite dans la Section 1.3.2. Après avoir extrait les aspects et identifié la polarité des segments de texte mentionnant ces aspects, l'approche la plus directe pour résumer ces informations consiste simplement à compter le nombre de segments positifs et négatifs (dans le cas où l'on suppose une polarité à deux niveaux) pour chaque aspect et à représenter ce volume d'opinions par exemple sous forme d'histogramme ou de boîtes à moustache. Cela permet ainsi de construire un résumé d'opinions visuel. Ce type d'approches a été adopté dans [Carenini *et al.*, 2006; Gamon *et al.*, 2005; Liu *et al.*, 2005]. La Figure 1.1 illustre le système Opinion Observer, développé par Liu *et al.* [2005]. L'avantage du résumé visuel est de faciliter la comparaison des points forts et des points faibles de différents produits ou marques entre lesquels un consommateur potentiel pourrait hésiter. Cependant, un résumé visuel basé sur le simple compte des opinions positives et

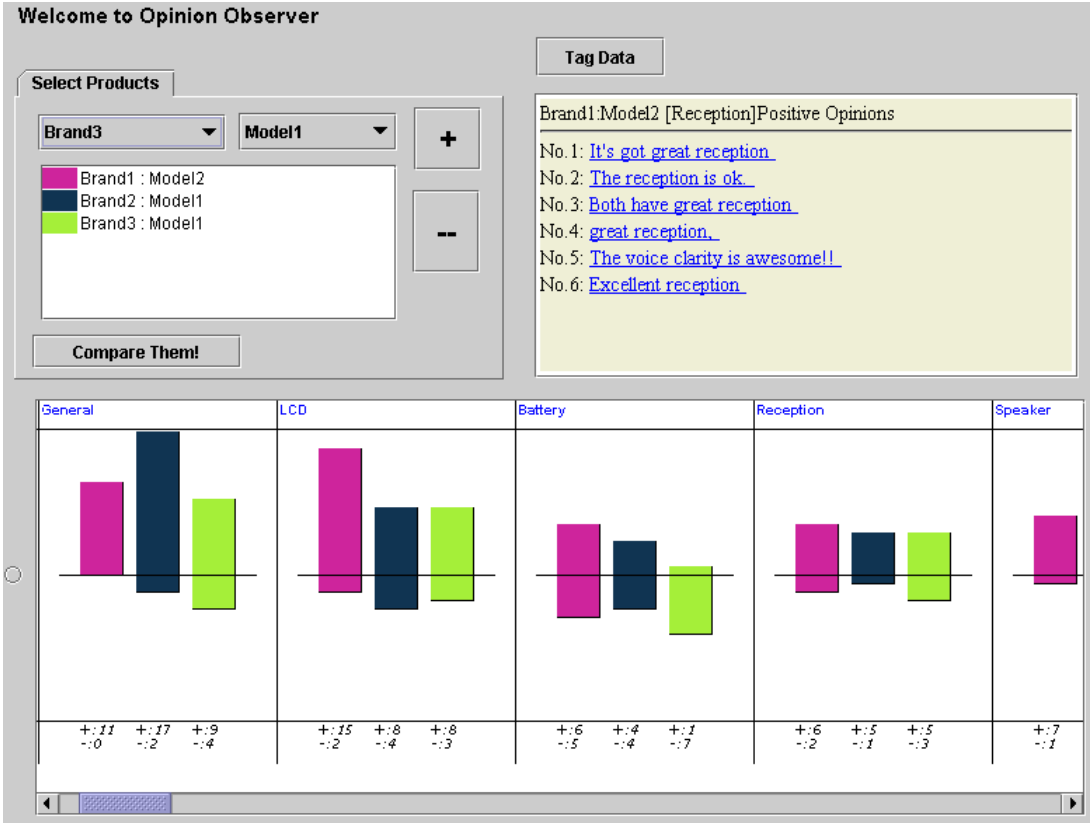


FIGURE 1.3 – Exemple de résumé d'opinions visuel basé sur les aspects. Capture d'écran du système Opinion Observer tirée de [Hu et Liu, 2004].

négatives pour chaque aspect peut fausser la perception de la cible. Par exemple, considérons un téléphone mobile A pour lequel 200 critiques mentionnent positivement la solidité de l'écran et un téléphone mobile B pour lequel 100 critiques mentionnent positivement la solidité de l'écran et 50 critiques notent la taille appréciable de l'écran. [15] Le compte des opinions pour l'aspect « écran » aura tendance à avantager A par rapport à B ($200 > 150$), alors que l'écran de B présente deux aspects positifs (solidité et taille) et l'écran de A n'en présente qu'un (solidité). Ainsi, il peut s'avérer pertinent de considérer les opinions non redondantes plutôt que le nombre total d'opinions.

Pour pallier cet inconvénient du résumé visuel, un résumé idéal d'opinions basé sur les aspects se présente de la manière suivante : pour chaque aspect de la cible, le résumé énumère les opinions positives et négatives *distinctes* – lorsque de telles opinions existent. La notion clé ici est donc l'identification de la représentativité des segments d'opinions – *in fine*, l'ensemble des segments représentatifs constitue le résumé. Étant donné que ce type de résumé est plus focalisé sur la sélection du contenu que sur la présentation, nous le désignerons dans le reste de cette section en tant que « résumé textuel ». Nous présentons un exemple construit manuellement de ce type de résumés dans le Tableau 1.1. Cette approche du résumé d'opinions

15. Nous supposons pour simplifier cet exemple qu'aucune opinion négative n'a été exprimée sur ces aspects de A et B.

TABLEAU 1.1 – Exemple de résumé d'opinions textuel basé sur les aspects pour un film de cinéma (imaginaire). Les opinions ont été construites manuellement.

Histoire	
+	Le suspens du film m'a tenu en haleine du début jusqu'à la fin, je suis resté scotché tout le long ! (15 occurrences)
+	J'ai apprécié que l'intrigue de cette adaptation soit fidèle au livre.
Jeu d'acteurs	
+	Brad Pitt était parfait dans son rôle !
+	On sentait une grande complicité entre les acteurs, ce qui se ressentait sur leur jeu.
−	Angelina Jolie m'a un peu déçu, on voyait clairement qu'elle faisait semblant de pleurer à la fin du film...
Décors	
−	Les décors du film étaient vraiment peu inspirés, à la lecture du livre original j'imaginais quelque chose de beaucoup plus grandiose.
−	Le film est censé se dérouler sur Mars mais l'immersion a été complètement cassée quand j'ai reconnu dans le film un endroit célèbre en Australie !
Effets spéciaux	
−	Les effets spéciaux étaient super cheap, je me croyais devant les anciens épisodes de la série Star Wars.

basé sur les aspects orientée vers la sélection du contenu central a été largement abordée dans la littérature [Fang *et al.*, 2014; Ganesan et Zhai, 2012; Ganesan *et al.*, 2010; Lu et Zhai, 2008; Mei *et al.*, 2007; Meng *et al.*, 2012; Wang *et al.*, 2014]. On distingue parmi ces travaux :

— la génération de résumés d'opinions extractifs, consistant à extraire de la collection de textes les phrases destinées à être intégrées au résumé : [Fang *et al.*, 2014; Lu et Zhai, 2008; Mei *et al.*, 2007; Meng *et al.*, 2012; Wang *et al.*, 2014] ;

— la génération de résumés d'opinions abstractifs, qui construit les phrases du résumé à partir des mots importants : [Ganesan et Zhai, 2012; Ganesan *et al.*, 2010].

Dans le reste de cette section, nous détaillons ces deux types de travaux.

1.4.1.1 Résumés extractifs

Dans leur approche extractive, Mei *et al.* [2007] s'appuient sur un modèle de thèmes et d'opinions, qui a été détaillé dans la Section 1.3.2. À partir de ce modèle, les phrases représentatives sont extraites en calculant pour chaque phrase et chaque thème (respectivement, chaque degré de polarité) la divergence de Kullback-Leibler entre la distribution empirique des mots dans la phrase et la distribution des mots pour le thème (respectivement, pour le niveau de polarité). Dans [Fang *et al.*, 2014; Meng *et al.*, 2012; Wang *et al.*, 2014], la génération de résumés extractifs est formulée par un problème d'optimisation combinatoire afin de favoriser l'intégration dans le résumé de segments informatifs. Fang *et al.* [2014] se basent sur

la décomposition duale (*dual decomposition*) et postulent que les phrases sont d'autant plus informatives qu'elles contiennent de mots d'opinion et d'aspect. Dans [Wang *et al.*, 2014], le problème d'optimisation est basé sur des fonctions sous-modulaires pour sélectionner successivement les phrases les plus informatives. L'avantage de ce type de fonctions est la garantie d'approcher l'optimum par un algorithme glouton.

Un problème différent de la génération de résumés d'opinions classique est abordé dans [Lu et Zhai, 2008] : aligner des critiques en ligne rédigées par des amateurs (c'est-à-dire des internautes « lambda ») avec des critiques réalisées par des experts, et identifier les opinions d'amateurs ne pouvant être alignées. Les auteurs ont nommé cette tâche « intégration d'opinion » (*opinion integration*). La motivation de ce travail provient de l'observation que les critiques d'experts sont souvent structurées (par exemple, par aspect) mais disponibles en petit nombre, tandis que les critiques d'amateurs sont disponibles en quantité massive mais sont peu structurées. L'alignement de ces deux types d'opinions peut alors faciliter leur exploitation. Pour résoudre ce problème, nommé « intégration d'opinion » (*opinion integration*) dans [Lu et Zhai, 2008], les auteurs proposent un modèle thématique semi-supervisé basé sur PLSA. Il est cependant à noter que le résumé proposé suivant cette méthode ne prend pas en compte la polarité des opinions.

1.4.1.2 Résumés abstractifs

La génération de résumés d'opinions peut être également considérée sous l'angle des résumés abstractifs, qui reposent sur la construction de phrases synthétiques plutôt que de les extraire directement des textes originaux. Ganesan et Zhai [2012] définissent un problème d'optimisation pour identifier ce que les auteurs nomment des « micropinions », c'est-à-dire des opinions clés exprimées par une séquence de 2 à 5 mots. Les micropinions sont donc obtenues en agrégeant des mots pour former une séquence qui n'apparaît pas nécessairement dans le corpus de textes, rapprochant ainsi l'approche proposée des méthodes abstractives. L'objectif du problème d'optimisation prend en compte à la fois la représentativité des micropinions ainsi que leur lisibilité. Une autre approche abstractive employée dans [Ganesan *et al.*, 2010] repose sur la construction d'un graphe de mots extraits des phrases d'opinion du corpus afin de capturer la redondance de ces phrases. Les phrases du résumé sont ensuite générées à partir des chemins grammaticalement valides du graphe passant par des mots centraux.

Dans cette section, nous avons décrit en détail les approches extractives et abstractives pour la génération de résumés d'opinions basés sur les aspects. Ces résumés ont l'avantage de permettre d'identifier aisément les opinions positives et négatives associées à chaque aspect d'une cible. Cependant, les opinions positives et négatives associées à un même aspect proviennent parfois de la différence de contextes ou de perspectives entre les auteurs de ces opinions. Les résumés d'opinions basés sur les aspects ne permettent pas d'identifier ces paires d'opinions contrastées portant sur un même aspect, qui peuvent toutefois être utiles au potentiel acheteur. À partir des ces paires, celui-ci peut prendre sa décision d'achat en se basant sur son propre contexte ou sa perspective. Nous étudions les travaux proposant de découvrir de telles paires dans la Section 1.4.2.

Tableau 1.2 – Exemple de résumé d'opinions contrasté pour un jeu vidéo (imaginaire). Les opinions ont été construites manuellement. Les indications entre parenthèses ont été ajoutées pour clarifier la perspective ou le contexte des différentes opinions.

Cible : JV 2000		
Aspect	Paire d'opinions contrastées	
	−	+
Durée de vie	Ayant déjà joué à beaucoup de jeux dans le même style, j'ai malheureusement trouvé que JV 2000 ne posait aucun challenge et je l'ai terminé en moins de 10 heures. *(perspective : expert)*	C'est la première fois que je joue à un jeu comme JV 2000 et je peux dire que j'en ai eu pour mon argent : il m'a fallu une bonne vingtaine d'heures pour arriver au bout. *(perspective : débutant)*
Doublage	La localisation en français était vraiment horrible et j'ai trouvé les voix des personnages complètement insipides. *(contexte : langue française)*	L'idée de faire appel à des acteurs américains célèbres pour doubler les personnages était excellente. *(contexte : langue anglaise)*

1.4.2 Résumés d'opinions contrastés

Afin de démontrer l'intérêt d'avoir à disposition des paires d'opinions contrastées, considérons l'exemple suivant. Un utilisateur d'un site de commerce en ligne hésite à acheter un jeu vidéo donné que nous nommerons « JV 2000 ». Pour prendre sa décision, il lit les critiques postées par les autres utilisateurs. Une critique, rédigée par un joueur très expérimenté dans ce type de jeu, mentionne que le jeu est trop court et trop facile. Une deuxième critique, écrite par un joueur plus novice, apprécie au contraire la durée de vie raisonnable du jeu. Une troisième critique, postée par un utilisateur français, reproche au doublage français des protagonistes du jeu d'être de mauvaise qualité. Enfin, une quatrième critique fait les louanges des voix originales (en anglais) des personnages. L'idéal pour le potentiel acheteur serait donc de se voir présenter les paires d'opinions contrastées issues des critiques 1 et 2 et des critiques 3 et 4. Le résumé résultant de cet alignement est illustré dans le Tableau 1.2. Ainsi, selon son propre contexte (ici, sa langue : français ou anglais) et sa perspective (ici, son niveau dans ce type de jeux : débutant ou expert), il pourra alors juger quelles opinions sont les plus pertinentes pour lui. Dans la littérature, ce type de résumés alignant les opinions similaires mais opposées selon les contextes ou perspectives est nommé « résumé d'opinions contrasté » (*contrastive opinion summary*).

Ce problème a été introduit dans [Kim et Zhai, 2009] et appliqué aux critiques en ligne. Le problème d'optimisation proposé vise à identifier des paires de phrases maximisant à la fois la représentativité de chaque phrase et le contraste entre les phrases de chaque paire. La similarité entre deux phrases est simplement basée sur les unigrammes qu'elles contiennent. Récemment, d'autres travaux ont abordé ce problème en l'appliquant aux sujets soulevant différents *points de vue* (ce que nous avions également nommé « opinions collectives » dans la

Tableau 1.3 – Exemple de résumé de points de vue sur l'avortement. Les opinions ont été construites manuellement. Les indications entre parenthèses ont été ajoutées pour expliciter la polarité des opinions vis-à-vis du thème.

Sujet : l'avortement		
Thème	Points de vue	
	Pro-choix	Pro-vie
Droit à disposer de son corps	Le corps est la propriété la plus fondamentale d'un individu. La décision de continuer une grossesse ou non doit revenir à la femme concernée. *(polarité : +)*	Dès le sixième jour de grossesse, l'embryon est physiologiquement différencié du corps de la mère. L'avortement dépasse donc les droits individuels de la mère. *(polarité : −)*
Religion	Les religions promeuvent un ensemble de principes archaïques et patriarcaux réduisant la femme à sa capacité à procréer. *(polarité : −)*	Un bébé, même sous la forme de fœtus, est une créature de Dieu et devrait bénéficier du droit inconditionnel à la vie. *(polarité : +)*

Section 1.1.1), tels que les sujets politiques [Guo *et al.*, 2015; Ren et de Rijke, 2015; Ren *et al.*, 2016]. Inspiré par la méthode d'intégration d'opinions proposée dans [Lu et Zhai, 2008], Guo *et al.* [2015] utilisent également un modèle thématique semi-supervisé basé sur PLSA afin d'aligner des opinions d'amateurs exprimées sur Twitter avec des opinions d'experts issues du site procon.org [16]. Ensuite, pour chaque opinion d'expert – qui peut être perçue comme un argument – une paire d'opinions contrastées (c'est-à-dire une opinion positive et une négative) rédigées par des amateurs est ajoutée au résumé.

Dans [Ren et de Rijke, 2015; Ren *et al.*, 2016], les auteurs se sont appuyés sur des modèles thématiques étendant LDA pour générer des résumés d'opinions contrastés. Le modèle proposé par Ren et de Rijke [2015] est basé sur le processus de restaurants chinois emboîtés (*nested chinese restaurant process*) afin de construire une hiérarchie de thèmes contenant les mots positifs, neutres et négatifs associés. À partir de cette hiérarchie, les phrases contrastées et couvrant des thèmes variés sont extraites par une méthode basée sur les processus ponctuels déterminantaux (*determinantal point process*). Le modèle thématique développé dans [Ren *et al.*, 2016] prend quant à lui en compte l'horodatage des documents afin de construire un résumé contrasté temporel sur les médias sociaux.

1.5 Conclusion

Nous avons vu dans ce chapitre que le processus de fouille d'opinions vis-à-vis d'une cible donnée se découpe en trois étapes. Tout d'abord, les textes subjectifs sont détectés en se

16. `http://www.procon.org/`

basant sur des lexiques d'opinions, sur les parties du discours, sur des classifieurs supervisés issus de l'apprentissage statistique, ou sur des règles et des motifs. Ensuite, la polarité des textes subjectifs est identifiée par des classifieurs supervisés ou des méthodes non supervisées telles que les modèles thématiques. La polarité peut être identifiée par rapport à la cible dans sa globalité ou bien en considérant chacun de ses aspects séparément. Enfin, les différentes opinions sont résumées, visuellement ou textuellement, de manière extractive ou abstractive, en les regroupant par aspect ou sous forme de paires contrastées. Au travers de ce processus, la polarité d'une opinion est toujours considérée comme unidimensionnelle et positionnée sur un axe négatif-positif. Cette polarité négative et positive est particulièrement adaptée au traitement d'opinions individuelles, mais l'est-elle pour les opinions collectives (ou points de vue) ?

Les travaux sur la génération de résumés d'opinions contrastés abordent cette notion de points de vue. Les auteurs de [Ren et de Rijke, 2015; Ren *et al.*, 2016] considèrent qu'un point de vue est un couple contenant un thème et une polarité. En réalité, suivant notre définition établie dans la Section 1.1.1, un point de vue est un « ensemble (d')idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc. », c'est-à-dire un *ensemble* de couples (thème, polarité), et non un unique couple. Par ailleurs, les différents couples d'un même point de vue peuvent associer des polarités différentes à des thèmes différents. Pour illustrer ce phénomène, considérons le sujet [17] de l'avortement. Sur le thème du droit des femmes à disposer de leur corps, les militants pro-choix exprimeront généralement un sentiment positif et les militants pro-vie un sentiment négatif. Cette polarité sera cependant inversée si les militants de chaque bord évoquent le thème de la religion. Dans ce cadre, les techniques de fouille d'opinions basées uniquement sur les polarités positive et négative sont insuffisantes pour traiter la notion de point de vue (par exemple, pour identifier les points de vue exprimés dans des documents). Elles ne permettent pas la construction d'un résumé de points de vue tel que celui illustré dans le Tableau 1.3, reprenant le sujet de l'avortement. Ainsi, nous détaillerons dans le Chapitre 2 les travaux abordant les points de vue en étendant la notion d'opinions au-delà du positif et du négatif.

17. Dans le contexte de la fouille de points de vue, « sujet » (*issue*) et « thème » (*topic* ou *theme*) supplantent respectivement « cible » et « aspect », qui sont plus couramment associés à la fouille d'opinions individuelles.

Chapitre 2

Fouille de points de vue

Sommaire

2.1 Introduction

Dans ce chapitre, nous proposons d'unifier les différents travaux en fouille d'opinions qui vont au delà de la conception positive et négative des opinions. Nous regroupons ces travaux sous le nom de fouille de points de vue, ou fouille d'opinions collectives, en nous basant sur la définition des points de vue établie dans la Section 1.1.1 du Chapitre 1. Un état de l'art sur une partie des travaux décrits dans ce chapitre est également disponible dans [Qiu, 2015].

Dans un premier temps, nous reviendrons sur la définition du point de vue et expliciterons les objectifs de la fouille de points de vue (Section 2.1.1). Ensuite, nous détaillerons les applications, difficultés et tâches additionnelles de la fouille de points de vue par rapport à la fouille d'opinions individuelles dans les Sections 2.1.2, 2.1.3 et 2.1.4, respectivement. Enfin, dans les Sections 2.2, 2.3 et 2.4, nous décrirons les travaux étudiant les points de vue à différents niveaux : au niveau du mot ou de la phrase (niveau microscopique), au niveau du document ou de l'utilisateur (niveau mésoscopique) et au niveau de la population et du sujet d'étude (niveau macroscopique), respectivement.

2.1.1 Définitions et objectifs

Rappelons la définition d'opinions collectives établie dans la Section 1.1.1 : une opinion collective est l'« ensemble des idées d'un groupe social sur les problèmes politiques, économiques, moraux, etc ». Afin de faciliter la distinction entre opinions individuelles et collectives, nous dénommerons par la suite les opinions collectives sous l'appellation « points de vue ». Ce choix est par ailleurs guidé par la terminologie utilisée dans la littérature (en anglais *viewpoints*, *points of view* ou *views*), par exemple dans [Ahmed et Xing, 2010; Cohen et Ruths, 2013; Conover *et al.*, 2011b; Garimella *et al.*, 2016, 2017; Graells-Garrido *et al.*, 2015; Guerra *et al.*, 2013; Jin *et al.*, 2016; Menini et Tonelli, 2016; Paul *et al.*, 2010; Qiu et Jiang, 2013; Trabelsi et Zaïane, 2014]. Il est à noter que les opinions collectives sont également parfois appelées « perspectives » (*perspectives*) [Fang *et al.*, 2012; Hardisty *et al.*, 2010; Lin et Hauptmann, 2006; Lin *et al.*, 2006, 2008; van der Zwaan *et al.*, 2016] et « positions » (*stances*, *sides* ou *positions*) [Augenstein *et al.*, 2016; Gottipati *et al.*, 2013; Johnson et Goldwasser, 2016; Qiu *et al.*, 2015, 2013b; Somasundaran et Wiebe, 2010].

Étant donnée la dimension sociale significative des points de vue, en tant que rattachement d'un individu à un groupe aux idées communes, les interactions sociales (par exemple sur les plateformes de réseaux sociaux telles que Twitter et Facebook) constituent des indices considérables pour l'étude des points de vue. Par conséquent, une distinction majeure entre la fouille d'opinions individuelles et la fouille de points de vue est que là où la première se base uniquement sur les données textuelles, la seconde peut en outre exploiter la structure de graphes décrite par les interactions sociales. Ainsi, la fouille de points de vue désigne le domaine d'étude des points de vue dans les textes et les réseaux sociaux. Nous conservons ici volontairement le terme d'« étude » qui, bien que vague, permet d'englober l'ensemble des tâches distinctes en fouille de points de vue, que nous définirons dans la Section 2.1.4.

2.1.2 Motivations et applications

Comme nous l'avions expliqué dans la Section 1.1.2, l'analyse de l'opinion publique constitue une tâche critique d'un point de vue économique et politique. La Section 1.1.2 avait décrit en profondeur l'intérêt économique des systèmes de fouille d'opinions individuelles pour les consommateurs et les fabricants. L'exemple choisi était l'application de tels systèmes à la génération de résumés d'opinions à partir de critiques en ligne. La présente section met au contraire l'accent sur les implications politiques et applications des systèmes de fouille de points de vue.

En plus de discuter de produits commerciaux dans les blogs et les réseaux sociaux, un autre sujet prédominant est la politique. Plusieurs sondages du *Pew Research Center*[1] en attestent [Duggan et Smith, 2016; Smith, 2014]. Un sondage conduit en 2014 [Smith, 2014] suggère que le nombre d'électeurs américains suivant des personnalités politiques sur les médias sociaux a plus que doublé entre 2010 et 2014 (6 % des électeurs en 2010 contre 16 % en

1. `http://www.pewresearch.org/`

2014). Ce pourcentage a continué de croître jusqu'en 2016, culminant alors à 25 % [Duggan et Smith, 2016]. Ce dernier sondage indique également qu'en 2016 environ un tiers (32 %) des utilisateurs de médias sociaux commentent, participent à des discussions ou postent du contenu en rapport avec le gouvernement et la politique. Par conséquent, les utilisateurs de médias sociaux ne sont pas passifs : ils sont moteurs dans la génération de contenu politisé.

Cet important volume de données d'opinions politiques semble susceptible de pouvoir complémenter (à défaut de supplanter [Kim *et al.*, 2014]) les opinions recueillies par les sondages classiques [O'Connor *et al.*, 2010]. Cette question a également été étudiée dans un rapport de l'association américaine pour l'étude de l'opinion publique (*American Association for Public Opinion Research*) [Murphy *et al.*, 2014]. Une autre application similaire est la prédiction de résultats d'élections présidentielles à partir des médias sociaux [Lampos *et al.*, 2013; Tsakalidis *et al.*, 2015; Tumasjan *et al.*, 2010]. Cette application a également été étudiée par Galam [2017] dans le cadre de l'élection présidentielle française de 2017 et a connu une forte couverture médiatique [2] [3] [4] [5]. Toutefois, plusieurs travaux par Gayo-Avello ont noté les limites des approches qui prétendent être capables de prédire des résultats d'élections à partir des médias sociaux [Gayo-Avello, 2011, 2012]. Les critiques formulées évoquent un biais de publication : seuls les résultats positifs – c'est-à-dire les prédictions « réussies » – sont publiés. Un biais de sélection et un biais démographique sont également mentionnés, liés respectivement à la méthode de sélection des données par le chercheur, et du manque de représentativité des différents groupes démographiques dans les médias sociaux par rapport à la population réelle. De plus, il est reproché à ces approches de se baser essentiellement sur des méthodes de fouille d'opinions basiques, ne prenant en compte que les opinions positives et négatives. Des systèmes de fouille de points de vue plus sophistiqués, allant au delà des opinions positives et négatives, pourraient constituer des approches plus robustes pour ce type d'applications.

Dans leur rapport de 2016 [Duggan et Smith, 2016], le *Pew Research Center* a également noté la forte propension des utilisateurs de médias sociaux à suivre les personnalités politiques qui partagent le même point de vue : 66 % des utilisateurs suivent des personnalités politiques qui partagent le même point de vue, alors que seulement 31 % (respectivement, 3 %) suivent des personnalités politiques aux points de vue variés (respectivement, opposés aux leurs). Ces chiffres montrent que les internautes – et, en particulier, les utilisateurs de médias sociaux – auront plus tendance à être exposés à du contenu soutenant leur propre point de vue. Ce phénomène, dénommé « chambre d'échos » par Sunstein [2009] et « bulle de filtres » par Pariser [2011], tirerait son origine dans l'homophilie (c'est-à-dire la propension à créer des liens avec les personnes partageant les mêmes idées) et la personnalisation de contenu. Il a récemment été l'objet de nombreuses études, par exemple dans le cadre des réseaux sociaux [Bakshy

2. https://theconversation.com/pourquoi-et-comment-marine-le-pen-peut-gagner-avec-moins-de-50-dintentions-de-vote-74994

3. http://www.lejdd.fr/Politique/Serge-Galam-le-physicien-qui-predit-la-victoire-de-Marine-Le-Pen-859634

4. https://www.sciencesetavenir.fr/politique/faut-il-croire-le-sociophysicien-qui-donne-marine-le-pen-gagnante-au-second-tour_112529

5. http://www.francetvinfo.fr/politique/emmanuel-macron/l-abstention-peut-faire-gagner-marine-le-pen_2118765.html

et al., 2015; Dunn *et al.*, 2015] ou de la consommation d'articles de presse en ligne [Flaxman *et al.*, 2016].

Bien qu'il ait été observé que les internautes ne sont pas nécessairement intéressés par du contenu purement antagoniste, une étude des motifs de consommation de presse en ligne sur un échantillon représentatif de la population américaine a suggéré que des sources équilibrées – qui reflètent des points de vue variés – sont néanmoins appréciées [Garrett et Stroud, 2014]. Ainsi, les systèmes de fouille de points de vue pourraient permettre aux utilisateurs d'accéder à des opinions variées (par exemple, sous forme de résumés de points de vue) et ainsi réduire l'effet des chambres d'échos et des bulles de filtre. Alternativement, un système de fouille de points de vue peut être mis en œuvre pour connecter ensemble des utilisateurs aux opinions variées [Garimella *et al.*, 2017]. Une autre application possible est la détection de fausses nouvelles (*fake news*) [Jin *et al.*, 2016] – un sujet qui a fait couler beaucoup d'encre depuis l'élection présidentielle américaine de 2016 [Allcott et Gentzkow, 2017] et qui a donné lieu à une compétition pour universitaires et industriels en 2017 [6].

Idéalement, un résumé de points de vue serait organisé sous forme de carte argumentative (*argument map*). Une telle représentation énumère les différents arguments (par exemple, regroupés par thème et par point de vue), ce qui s'avère clé pour décrire les discussions résultant d'un débat. Il a également été montré, dans un cadre éducatif, que l'utilisation de cartes argumentatives aide au développement de la pensée critique [Twardy, 2004]. Un exemple de carte argumentative est illustré dans la Figure 2.1 et sa version annotée est fournie dans la Figure 2.2. Cette carte, mentionnant les arguments en soutien et en opposition à l'utilisation du gaz de schiste en Europe, a été mis a disposition par TNO [7] et The Argumentation Factory [8]. Notons que cette carte a été construite manuellement, par des experts. Il serait par conséquent précieux de disposer de systèmes capables de constituer automatiquement une carte argumentative pour un sujet donné, en délimitant tout d'abord les thèmes (comme savent le faire les modèles thématiques) puis en identifiant les arguments des différents points de vue vis-à-vis de ces thèmes. Nous décrivons dans la section suivante les difficultés que présente la construction d'un tel système et expliquons dans quelle mesure les méthodes de fouille d'opinions individuelles, décrites dans le Chapitre 1, sont insuffisantes pour répondre à ce besoin.

2.1.3 Difficultés et spécificités par rapport à la fouille d'opinions individuelles

Similairement aux systèmes de fouilles d'opinions individuelles, les systèmes de fouilles de points de vue sont confrontés à une difficulté d'ordre lexical : les mots utilisés pour exprimer différentes opinions (individuelles ou collectives) associées à une cible donnée sont généralement très semblables, puisqu'ils se rapportent aux mêmes thèmes. L'identification des différentes opinions est par conséquent difficile si elle est uniquement basée sur l'occur-

6. `http://www.fakenewschallenge.org/`
7. `https://www.tno.nl/en/`
8. `https://www.argumentenfabriek.nl/`

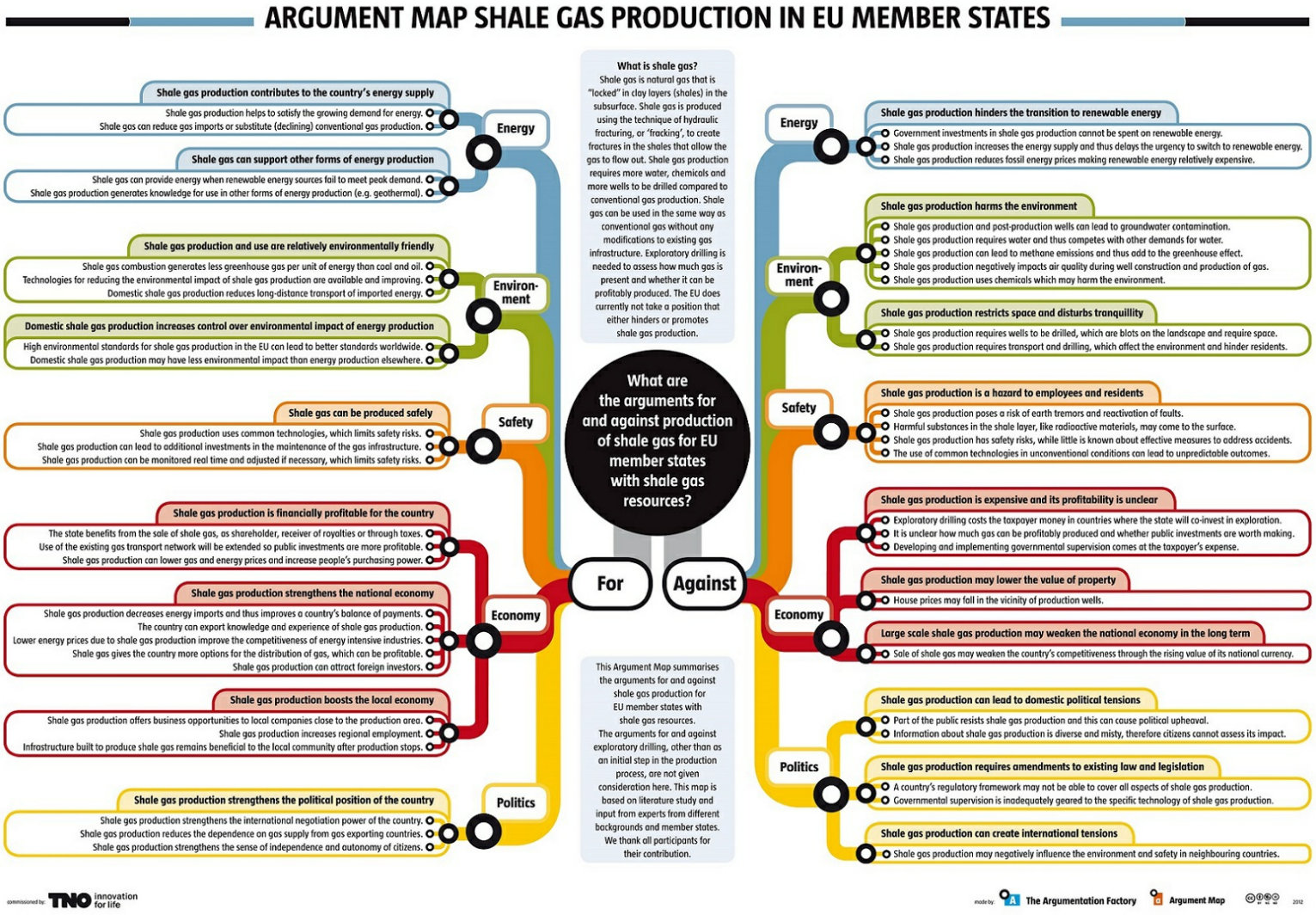


FIGURE 2.1 – Exemple de carte argumentative sur le gaz de schiste mise à disposition par TNO (`https://www.tno.nl/en/`) et The Argumentation Factory (`https://www.argumentenfabriek.nl/`).

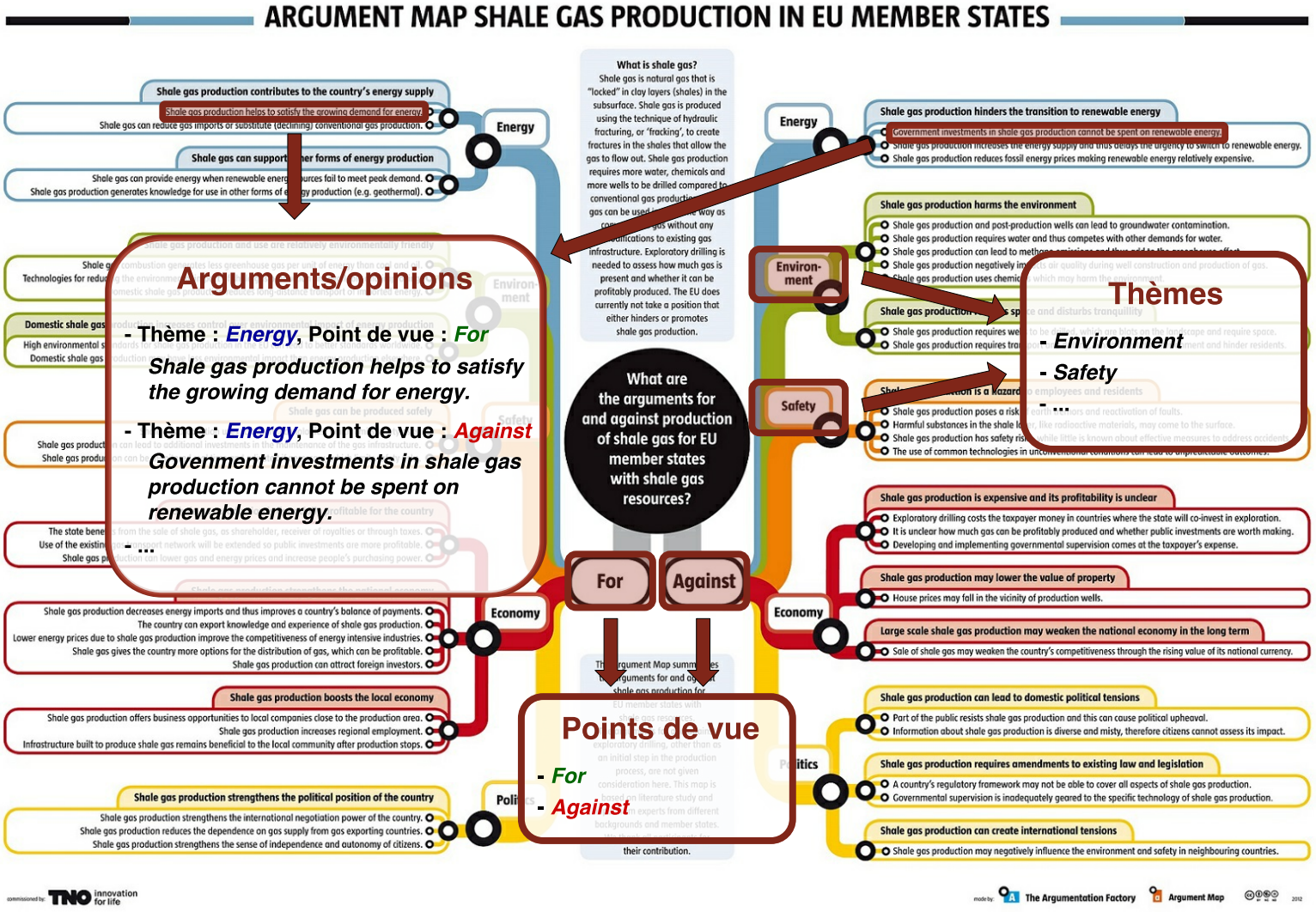


FIGURE 2.2 – Annotation de la carte argumentative sur le gaz de schiste. Les moitiés gauche et droite illustrent les deux points de yue (*for* et *against*, respectivement). Les branches issues des points de vue correspondent aux thèmes (par exemple, *environment*, *safety*). Les feuilles contiennent quant à elle les différents arguments ou opinions associés à chaque thème et propres à un point de vue.

rence des mots dans les textes. Pour pallier ce problème, comme nous l'avons mentionné dans le Chapitre 1, de nombreuses approches de fouille d'opinions individuelles se basent sur des lexiques d'opinions prédéfinis, recensant un ensemble de mots généraux de polarité positive et négative.

Cependant, si l'on étudie les points de vue – par exemple, dans l'optique d'identifier selon quel point de vue un document a été rédigé, ces lexiques perdent leur utilité : il ne suffit plus de savoir qu'un document contient des mots positifs ou des mots négatifs pour déterminer un point de vue. Par ailleurs, l'expression du point de vue est souvent plus subtile que l'expression d'une opinion positive ou négative. Pour rendre compte de cette subtilité, considérons les exemples suivants :

1. Pendant la guerre des Six Jours, Israël a *occupé les territoires palestiniens* de la bande de Gaza.
2. Pendant la guerre des Six Jours, Israël a *aménagé des implantations* dans la bande de Gaza.

Bien que ces deux exemples relatent exactement le même fait, ils sont rédigés selon deux points de vue distincts : pro-palestinien pour le premier et pro-israélien pour le second. Les expressions qui révèlent ces points de vue sont « occupé les territoires palestiniens » et « aménagé des implantations ». Sans ces indices subtils, les points de vue de ces deux exemples ne pourraient être identifiés. Il pourrait être alors tentant de construire un lexique spécifique aux points de vue du sujet étudié, mais ce lexique se révélerait inapplicable à d'autres sujets. Par exemple, un lexique reflétant les points de vue républicain et démocrate dans le contexte de la politique aux États-Unis ne pourra clairement pas être réutilisé pour étudier les arguments des partisans et opposants de l'utilisation du gaz de schiste en Europe. Par conséquent, les méthodes à base de lexiques présentent un intérêt limité en fouille de points de vue.

Si un système de fouille de points de vue ne peut s'appuyer sur des lexiques, il peut néanmoins bénéficier d'indicateurs issus des médias sociaux (par exemple, le *retweet*, le *follow* ou l'utilisation de *hashtags* sur Twitter) lorsque les textes étudiés en sont issus. Par exemple, l'étude du phénomène d'homophilie sur les réseaux sociaux et en particulier dans le cas de la polarisation politique a établi que certaines interactions sociales comme le *retweet* constituent des signaux forts de l'approbation par un utilisateur du contenu posté par un autre utilisateur [Conover *et al.*, 2011b]. Le défi réside toutefois dans l'exploitation de ces indicateurs qui, bien que potentiellement utiles pour la fouille de points de vue, demeurent malgré tout des preuves implicites.

2.1.4 Scénarios de fouille de points de vue

La fouille de points de vue peut être effectuée à différents niveaux de granularité du texte, définissant ainsi plusieurs scénarios et tâches. Nous distinguons les trois niveaux suivants :

— le *niveau microscopique*, c'est-à-dire le niveau des mots, des groupes de mots, des phrases ou des documents « courts » (de la taille d'une phrase) ;

— le *niveau mésoscopique*, c'est-à-dire le niveau des documents « longs » (composés de plusieurs phrases ou paragraphes) et le niveau des utilisateurs de médias sociaux, considérés comme l'agrégat de l'ensemble des textes qu'ils ont postés ;

— le *niveau macroscopique*, c'est-à-dire le niveau des populations ou des sujets/thématiques.

Le niveau microscopique, que nous détaillerons dans la Section 2.2, comprend la détection d'expressions de contention et d'argumentation – consistant respectivement à identifier si une expression témoigne d'un désaccord ou formule un argument – ainsi que la classification de points de vue dans les documents courts. Au niveau mésoscopique, nous étudierons dans la Section 2.3 les travaux portant sur l'identification des points de vue dans les documents longs et des points de vue associés aux utilisateurs de médias sociaux. Enfin, les travaux de fouille de points de vue positionnés au niveau macroscopique, portant sur l'analyse comparative de points de vue à travers différentes populations et sur la détection de sujets de controverse, seront décrits dans la Section 2.4.

2.2 Fouille au niveau microscopique : mots et phrases

Comme il l'a été illustré dans l'exemple de la Section 2.1.3 sur la nuance de point de vue induite par les expressions « occupé les territoires palestiniens » et « aménagé des implantations », le simple choix des mots peut avoir un impact important – quoique subtil – sur le point de vue perçu. Pour cette raison, de nombreux travaux ont étudié le point de vue au niveau du mot, du groupe de mots et de la phrase, ou dans les documents courts de la taille d'une phrase (tel qu'un *tweet*, contenant au maximum 140 caractères). Nous qualifions ce niveau de « microscopique » car il se focalise sur l'unité atomique du langage – le mot – et la composition d'un nombre limité de mots.

La littérature distingue deux types de travaux sur la fouille de points de vue au niveau microscopique :

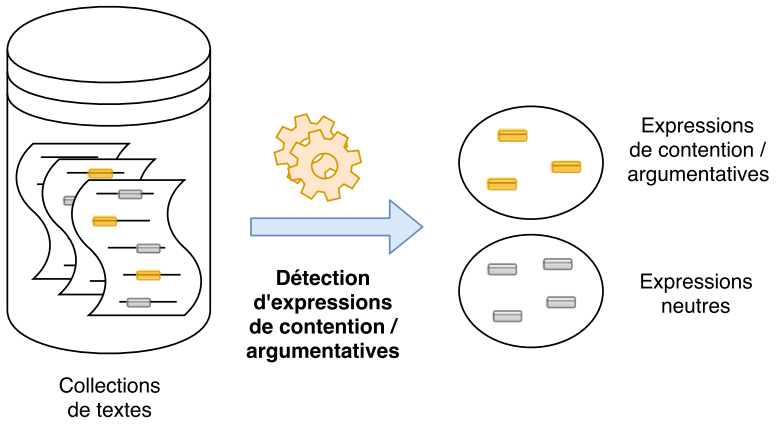
— les travaux sur la détection d'expressions d'argumentation et de contention [Cabrio et Villata, 2012; Galley *et al.*, 2004; Levy *et al.*, 2014; Lippi et Torroni, 2015; Menini et Tonelli, 2016; Mukherjee et Liu, 2012, 2013; Trabelsi et Zaïane, 2014, 2015, 2016] (Figure 2.3a) ;

— les travaux sur la classification de points de vue dans les documents courts [Augenstein *et al.*, 2016; Awadallah *et al.*, 2012; Iyyer *et al.*, 2014; Pennacchiotti et Popescu, 2011; Rao et Spasojevic, 2016; Somasundaran et Wiebe, 2010] (Figure 2.3b).

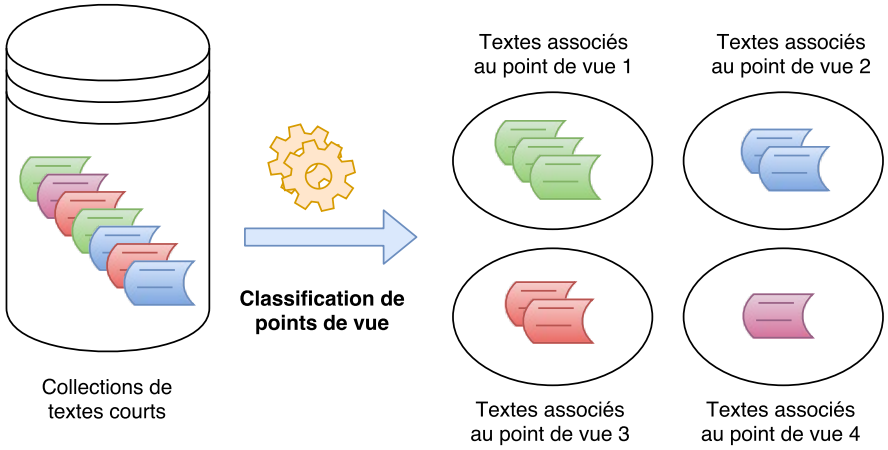
Dans le reste de cette section, nous détaillons ces différents travaux.

2.2.1 Détection d'expressions d'argumentation et de contention

L'expression d'un point de vue sous forme verbale, par exemple dans le cadre d'un débat, peut conduire à formuler des arguments pour soutenir une position ou à témoigner son accord ou son désaccord vis-à-vis d'arguments exprimés par des tiers. Nous dénotons les expressions



(a) Détection d'expressions d'argumentation et de contention. L'objectif est de distinguer les fragments de texte exprimant de la contention ou un argument des fragments de texte neutres.



(b) Classification de points de vue dans les documents courts. L'objectif est de classer les textes courts en fonction du point de vue qu'ils expriment.

FIGURE 2.3 – Tâches de fouille de points de vue au niveau microscopique.

d'accord ou de désaccord sous l'appellation d'« expressions de contention [9] » (*contention expressions*).

La détection d'arguments est rattachée au domaine de la fouille d'arguments (*argument mining*), qui a récemment été l'objet de plusieurs ateliers (*workshops*) dans la communauté du traitement automatique du langage naturel de 2014 à 2017 [Cardie, 2015; Green *et al.*, 2014; Reed, 2016]. Plusieurs travaux ont abordé la détection d'arguments sous la forme d'un problème de classification supervisée, par exemple par régression logistique [Levy *et al.*, 2014] ou par machine à vecteurs de support [Lippi et Torroni, 2015], en cherchant à répondre à la question suivante : est-ce qu'une phrase donnée contient une affirmation (*claim*) ? Levy *et al.* [2014] considèrent les affirmations dépendantes d'un contexte connu (le sujet à propos duquel sont formulés les arguments), tandis que Lippi et Torroni [2015] supposent que le

9. Dans ce contexte, la contention est définie par le Centre national de ressources textuelles et lexicales comme le « débat, (la) querelle, (la) situation contentieuse » (`http://www.cnrtl.fr/definition/contention`).

contexte est inconnu. Dans [Cabrio et Villata, 2012], la détection des arguments et de leur relation (soutien ou contradiction) s'appuie sur une méthode supervisée d'implication textuelle (*textual entailment*).

Le problème similaire de détection d'expressions de contention a également été largement abordé dans la littérature [Galley *et al.*, 2004; Menini et Tonelli, 2016; Mukherjee et Liu, 2012, 2013; Trabelsi et Zaïane, 2014, 2015, 2016]. Ce problème a également été étudié sous l'angle de la classification supervisée, en appliquant un classifieur à maximum d'entropie [Galley *et al.*, 2004] et une machine à vecteurs de support [Menini et Tonelli, 2016]. Cependant, d'autres travaux ont proposé d'exploiter des modèles thématiques [Mukherjee et Liu, 2012, 2013; Trabelsi et Zaïane, 2014, 2015, 2016] afin de s'affranchir en partie ou totalement de la nécessité de disposer de données annotées. Mukherjee et Liu [2012, 2013] ont ainsi adopté une approche semi-supervisée pour identifier les expressions de contention sur les forums de débat. Elle se base sur un modèle thématique qui intègre un *prior* utilisant la sortie d'un classifieur à maximum d'entropie préalablement entraîné. L'avantage de cette méthode est que même si le classifieur est entraîné sur un nombre limité de données annotées (spécifiant si une expression est de contention ou non), le mécanisme de *prior* du modèle thématique permettra de guider efficacement l'identification de la nature (de contention ou non) des expressions non annotées. Trabelsi et Zaïane [2014, 2015, 2016] ont au contraire postulé que les expressions d'accord et de désaccord sont principalement limitées aux forums, en raison de l'organisation de ceux-ci en fils de discussion (*discussion threads*). Par conséquent, ils ont généralisé la notion d'expression de contention en se basant sur les points de vue. Ainsi, le modèle thématique JTV (*joint topic viewpoint*) proposé dans [Trabelsi et Zaïane, 2014, 2015, 2016] modélise conjointement les thèmes et les points de vue spécifiques aux thèmes. Après avoir appliqué JTV à une collection de textes, les expressions de contention sont obtenues en appliquant un algorithme similaire aux k-moyennes (k-*means*) afin de regrouper les expressions associées au même point de vue.

2.2.2 Classification de points de vue dans les documents courts

Plutôt que d'extraire les expressions spécifiques à un point de vue, d'autres travaux ont cherché à classer des phrases ou des documents courts tels que des *tweets* selon le point de vue qu'ils reflètent. Ce problème a fait l'objet d'une tâche dans la campagne d'évaluation SemEval en 2016 [Mohammad *et al.*, 2016]. Cette tâche, intitulée « prédiction de positions » (*stance prediction*), comprend à la fois une sous-tâche supervisée et une sous-tâche faiblement supervisée. La première a pour objectif de prédire le point de vue exprimé dans des *tweets* sur 5 sujets différents (par exemple, « la légalisation de l'avortement »), en ayant à disposition des *tweets* annotées pour ces 5 sujets. La seconde sous-tâche, faiblement supervisée, consiste à identifier le point de vue de *tweets* portant sur un 6$^{\text{e}}$ sujet inédit en exploitant les données annotées des 5 premiers sujets ainsi que des données non annotées portant sur ce 6$^{\text{e}}$ sujet.

Afin de classer les documents courts ou phrases en fonction du point de vue qu'ils expriment, plusieurs travaux ont employé des méthodes basées sur les réseaux de neurones [Augenstein *et al.*, 2016; Iyyer *et al.*, 2014; Rao et Spasojevic, 2016]. Iyyer *et al.* [2014] ont proposé

une approche supervisée utilisant un réseau de neurones récursif (*recursive neural network*) pour modéliser le phénomène de composition sémantique du langage. Dans [Augenstein *et al.*, 2016; Rao et Spasojevic, 2016], la dépendance entre mots successifs est intégrée par le biais d'un réseau de neurones récurrent (*recurrent neural network*) de type LSTM (*long short-term memory*). La méthode développée par Rao et Spasojevic [2016] est supervisée, tandis que celle de Augenstein *et al.* [2016] est seulement faiblement supervisée, appliquée à la seconde sous-tâche de prédiction de positions de SemEval 2016. D'autres méthodes supervisées ont également été employées dans la littérature pour cette tâche, telles que les machines à vecteurs de support [Somasundaran et Wiebe, 2010] ou la méthode des k plus proches voisins (*k-nearest neighbors*) [Awadallah *et al.*, 2012]. Les traits utilisés par ces approches sont de nature lexicale : ils dépendent de l'occurrence des mots dans les documents courts ou phrases étudiées.

Les différents travaux que nous avons décrits dans la Section 2.2 abordent la fouille de points de vue au niveau microscopique, en considérant seulement des courts fragments de texte. Plutôt que considérer ces fragments de manière isolée, l'alternative consiste à se positionner au niveau que nous nommons « mésoscopique » en étudiant les documents dans leur intégralité (au lieu des phrases qui le composent), ou l'ensemble des *posts* rédigés par un utilisateur de médias sociaux (au lieu des *posts* de manière indépendante). L'avantage de cette approche est de pouvoir exploiter plus efficacement le phénomène de co-occurrence des mots dans les textes. Nous détaillons dans la Section 2.3 les travaux abordant la fouille de points de vue au niveau mésoscopique.

2.3 Fouille au niveau mésoscopique : documents longs et utilisateurs

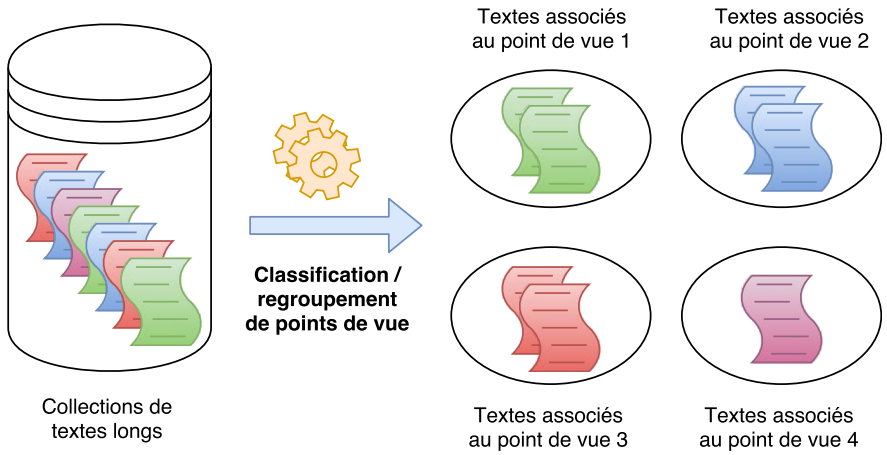
De manière comparable aux travaux au niveau microscopique décrits dans la Section 2.2.2, les travaux de fouille de points de vue positionnés au niveau mésoscopique dans la littérature ont également pour objectif d'identifier le point de vue d'un texte ou d'un agrégat de textes. Cependant, ces travaux sont distincts par leur méthodologie. Les travaux mentionnés dans la Section 2.2.2 nécessitent tous un certain degré de supervision pour classer les documents courts et n'utilisent pas les interactions sociales entre utilisateurs dans le cas des *posts* dans les médias sociaux. À l'inverse, les travaux que nous détaillons dans cette section sont moins dépendants des données annotées ou exploitent les interactions sociales. Nous distinguons deux tâches de fouille de points de vue au niveau mésoscopique :

— l'identification de points de vue dans les documents longs de type essais [Hardisty *et al.*, 2010; Lin *et al.*, 2006, 2008; Paul et Girju, 2010; Paul *et al.*, 2010] et textes législatifs [Akoglu, 2014; Gerrish et Blei, 2011, 2012; Nguyen *et al.*, 2015; Poole et Rosenthal, 1985] (Figure 2.4a) ;

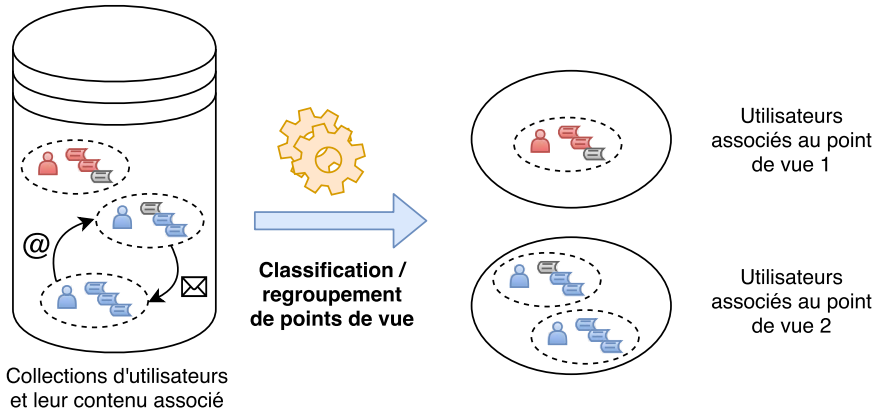
— l'identification du point de vue des utilisateurs dans les médias sociaux par des approches supervisées [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Conover *et al.*, 2011a; Fang *et al.*, 2015; Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011; Zhou *et al.*, 2015] et

semi-supervisées ou non supervisées [Abu-Jbara *et al.*, 2012; Bamman et Smith, 2015; Barberá, 2015; Gottipati *et al.*, 2014; Joshi *et al.*, 2016; Liu *et al.*, 2016, 2014; Qiu et Jiang, 2013; Qiu *et al.*, 2015, 2013a; Sachan *et al.*, 2012, 2014] (Figure 2.4b).

Les travaux portant sur ces deux tâches sont décrits dans les Sections 2.3.1 et 2.3.2, respectivement.



(a) Identification de points de vue dans les documents longs. L'objectif est de classer ou regrouper les textes longs en fonction du point de vue qu'ils expriment.



(b) Identification du point de vue des utilisateurs dans les médias sociaux. L'objectif est de classer ou regrouper les utilisateurs de médias sociaux en fonction du point de vue qu'ils expriment à partir du contenu textuel qu'ils ont généré et de leurs interactions sociales.

FIGURE 2.4 – Tâches de fouille de points de vue au niveau mésoscopique.

2.3.1 Identification de points de vue dans les documents longs

Les travaux portant sur l'identification de points de vue dans les documents longs ont exploré deux types de documents : les essais (Section 2.3.1.1) et les textes législatifs (Section 2.3.1.2).

2.3.1.1 Points de vue dans les essais

À notre connaissance, la tâche d'identification de points de vue dans les documents longs a été pour la première fois étudiée par Lin *et al.* [2006]. La tâche a alors été abordée comme un problème de classification supervisée et les auteurs ont comparé classifieurs bayésiens naïfs et machines à vecteurs de support pour résoudre ce problème. La contribution majeure de [Lin *et al.*, 2006] a été d'introduire auprès de la communauté scientifique la collection *Bitterlemons*. Bitterlemons est à l'origine un *e-zine*[10] bi-hebdomadaire publiant des essais rédigés par des auteurs israéliens et palestiniens sur le conflit israelo-palestinien ainsi que les questions géopolitiques s'y rapportant. La collection compilée par Lin *et al.* [2006] comprend en particulier 297 essais écrits par des auteurs israéliens et 297 écrits par des auteurs palestiniens. La taille moyenne des essais est d'environ 800 mots. Cette collection est donc idéale pour évaluer les approches visant à identifier les points de vue exprimés dans les documents longs.

À la suite de l'article fondateur de Lin *et al.* [2006], d'autres chercheurs se sont à leur tour intéressés à l'identification de points de vue dans les essais [Ahmed et Xing, 2010; Hardisty *et al.*, 2010; Lin *et al.*, 2008; Paul et Girju, 2010; Paul *et al.*, 2010] Hardisty *et al.* [2010] ont développé une extension du classifieur bayésien naïf proposé dans [Lin *et al.*, 2006] en se basant sur le mécanisme de grammaire adaptative *adaptor grammar*, ce qui permet au classifieur d'échapper aux restrictions imposées par l'hypothèse du sac de mots. D'autres travaux ont défini des modèles thématiques (étendant le modèle LDA) spécifiques à cette tâche, afin de limiter la dépendance aux annotations [Ahmed et Xing, 2010; Lin *et al.*, 2008; Paul et Girju, 2010; Paul *et al.*, 2010]. Le modèle thématique mview-LDA (*multi-view LDA*), décrit dans [Ahmed et Xing, 2010], est décliné en une version supervisée et une version semi-supervisée. En plus de modéliser le thème au niveau du mot, comme le fait LDA, mview-LDA ajoute une variable discrète dénotant le point de vue au niveau du document. Ces variables de point de vue sont considérées comme totalement observées dans le cas supervisé et partiellement observées dans le cas semi-supervisé. Le modèle mview-LDA a également la spécificité de permettre à un mot d'être tiré suivant 3 types de distributions : les distributions dépendant uniquement du point de vue, les distributions dépendant uniquement du thème et les distributions dépendant à la fois du thème et du point de vue. Le choix de la distribution selon laquelle un mot est tiré est déterminé par des variables de commutation *switch variables* intégrées au modèle.

On retrouve un mécanisme similaire dans le modèle TAM (*topic-aspect model*), proposé par Paul et Girju [2010]. TAM modélise conjointement et sans aucune supervision les thèmes et

10. `http://www.bitterlemons.net/`

les points de vue [11], qui sont ainsi considérés comme des variables latentes et sont positionnés au niveau du mot. Par le même mécanisme de variables de commutation que celui de mview-LDA, TAM permet à un mot d'être tiré suivant 4 distributions différentes : une distribution dépendant uniquement du point de vue, une distribution dépendant uniquement du thème, une distribution dépendant à la fois du thème et du point de vue (de manière identique à mview-LDA), ainsi qu'une distribution de fond (*background distribution*), indépendante du thème et du point de vue, destinée aux mots vides de sens ou non informatifs. Par la suite, TAM a été intégré à un système de génération de résumés de points de vue [Paul *et al.*, 2010].

2.3.1.2 Points de vue dans les textes législatifs

D'autres travaux ont abordé les points de vue exprimés dans un type différent de documents longs : les textes législatifs. L'étude du point de vue (également nommé « idéologie » ou *ideal point* dans ce contexte) des législateurs, tels que les membres du Congrès dans le système législatif américain, est l'objet d'une longue tradition en science politique. Par exemple, la famille de méthodes statistiques de positionnement multidimensionnel (*multidimensional scaling*) NOMINATE [Poole et Rosenthal, 1985] a été proposée pour visualiser la position idéologique des législateurs à partir de leurs votes sur des textes de loi (*bills*). De manière comparable, Akoglu [2014] a plus récemment développé un algorithme de propagation de polarité signée (*signed polarity propagation algorithm*) en modélisant les votes des législateurs sous la forme d'un graphe biparti signé.

L'inconvénient de ces méthodes est toutefois d'exploiter uniquement les données de vote et non les données textuelles des documents législatifs. Afin de pallier cette limite, plusieurs modèles thématiques étendant LDA ont été proposés pour modéliser conjointement les textes de loi et les votes qui leur sont associés [Gerrish et Blei, 2011, 2012; Nguyen *et al.*, 2015]. Comparé à LDA où les seules variables observées sont les mots apparaissant dans un document, ces modèles thématiques ajoutent des variables observées correspondant aux votes délivrés par chaque législateur sur chaque texte de loi. De plus, dans ce contexte, les points de vue sont modélisés sous la forme de variables continues et distribuées selon une loi normale, contrairement aux modèles thématiques mentionnés dans la Section 2.3.1.1. L'utilisation de variables continues rend ainsi possible la modélisation de points de vue nuancés : certains législateurs sont en effet plus « extrêmes » que d'autres dans leurs positions.

Le modèle IPTM (*ideal point topic model*) proposé par Gerrish et Blei [2011] introduit au niveau du document des variables de réponse (*response variables*) dépendantes du thème, à la manière du modèle LDA supervisé (*supervised LDA*). La différence majeure avec LDA supervisé est que dans IPTM, les variables de réponse sont latentes (c'est-à-dire, non observées), ce qui rend le modèle non supervisé. Ces variables de réponse font le lien entre les thèmes détectés dans un document et les votes sur ce document, et modulent l'impact de ces dernières sur le point de vue. Cela permet de modéliser l'intuition selon laquelle certaines lois

11. Les aspects considérés dans TAM sont proches de la notion de point de vue et ne doivent pas être confondus avec les aspects mentionnés en fouille d'opinions individuelles. Ces derniers sont nommés thèmes en fouille de points de vue.

sont plus révélatrices du point de vue des législateurs que d'autres. Par exemple, une loi sur le thème de l'avortement sera généralement plus clivante aux États Unis qu'une loi sur le thème de l'emploi. Une version alternative d'IPTM inspirée du modèle LDA étiqueté (*labeled LDA*) a été définie dans [Gerrish et Blei, 2012]. Le principe de cette méthode est d'introduire une faible supervision en associant les thèmes à des mots clés identifiés *a priori*, afin de guider la découverte des thèmes. Nguyen *et al.* [2015] ont par la suite proposé HIPTM (*hierarchical ideal point topic model*) qui structure les thèmes selon deux niveaux afin de définir une hiérarchie. Les thèmes du premier niveau sont alignés avec des sujets prédéfinis (par exemple, la santé, l'agriculture, l'emploi). Les thèmes du deuxième niveau sont quant à eux découverts de manière non supervisée et leur nombre n'est pas fixé mais appris à partir des données. Ce type de modèles, permettant l'apprentissage du nombre de thèmes lors de l'inférence, est désigné comme « non paramétrique » dans la littérature. Cela est rendu possible en faisant appel aux processus de Dirichlet (*Dirichlet processes*) plutôt qu'aux distributions de Dirichlet classiquement utilisées dans LDA.

Les travaux mentionnés dans cette section se focalisent sur les points de vue des législateurs, qui ont la particularité d'être relativement tranchés et explicites. En effet, chaque législateur est rattaché à un parti politique, qui lui-même se positionne fermement dans le spectre politique. Par conséquent, lorsqu'un législateur exprime son vote sur un texte de loi, il agit généralement en corrélation avec les consignes de son parti. Au contraire, l'identification du point de vue de personnes « lambda », telles que les utilisateurs de média sociaux, est un problème plus complexe : leur point de vue est exprimé de manière moins claire et parfois noyé dans le bruit caractéristique des médias sociaux. Nous détaillons les travaux qui ont abordé cette question dans la Section 2.3.2.

2.3.2 Identification du point de vue des utilisateurs de médias sociaux

La tâche d'identification du point de vue des utilisateurs dans les médias sociaux a été l'objet d'un grand nombre de travaux, motivés par l'impact croissant des médias sociaux sur la société et en particulier sur la politique [Duggan et Smith, 2016; Smith, 2014]. Cette tâche se distingue de l'identification des points de vue dans les documents longs tels que les essais et les textes législatifs dans la mesure où les messages postés sur les médias sociaux sont généralement courts et informels. De plus, une particularité des médias sociaux est de permettre à ses utilisateurs d'interagir (par exemple, grâce aux mécanismes de *follow*, de *retweet* et de *mention* sur Twitter). Le problème de l'identification du point de vue des utilisateurs dans les médias sociaux a été abordé dans la littérature suivant différents niveaux de supervision. Nous détaillons ainsi dans le reste de cette section :

— les approches supervisées [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Conover *et al.*, 2011a; Fang *et al.*, 2015; Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011; Zhou *et al.*, 2015] ;

— les approches semi-supervisées et non supervisées [Abu-Jbara *et al.*, 2012; Bamman et Smith, 2015; Barberá, 2015; Gottipati *et al.*, 2014; Joshi *et al.*, 2016; Liu *et al.*, 2016, 2014; Qiu et Jiang, 2013; Qiu *et al.*, 2015, 2013a; Sachan *et al.*, 2012, 2014].

2.3.2.1 Approches supervisées

Les approches supervisées proposées pour identifier le point de vue des utilisateurs dans les médias sociaux [12] se basent généralement sur des classifieurs tels que les machines à vecteurs du support [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Conover *et al.*, 2011a; Magdy *et al.*, 2016; Zhou *et al.*, 2015], des classifieurs bayésiens naïfs [Fang *et al.*, 2015] ou des arbres de décision [Pennacchiotti et Popescu, 2011]. Bien que ces travaux se rapprochent en apparence de ceux portant sur la classification de points de vue dans les documents courts (décrits dans la Section 2.2.2), la différence majeure est qu'au niveau des utilisateurs, les interactions sociales peuvent être prises en compte. Les interactions sociales constituent de précieux indices pour identifier les points de vue étant donné que les utilisateurs aux opinions similaires sont plus susceptibles de créer des liens [Conover *et al.*, 2011b] – une observation en accord avec le phénomène d'homophilie.

Les classifieurs proposés pour prédire le point de vue des utilisateurs sur Twitter combinent ainsi des traits exploitant différents indices :

— le contenu des tweets (mots [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Conover *et al.*, 2011a; Fang *et al.*, 2015; Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011; Zhou *et al.*, 2015], *hastags* [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Conover *et al.*, 2011a; Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011]) ;

— les informations indiquées dans le profil des utilisateurs [Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011] ;

— les interactions entre utilisateurs (*follow* [Pennacchiotti et Popescu, 2011; Zhou *et al.*, 2015], *mention* [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Magdy *et al.*, 2016], *reply* [Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011], *retweet* [Al Zamal *et al.*, 2012; Cohen et Ruths, 2013; Magdy *et al.*, 2016; Pennacchiotti et Popescu, 2011]).

En outre, plusieurs approches ont au préalable appliqué un modèle thématique tel que LDA [Fang *et al.*, 2015; Pennacchiotti et Popescu, 2011] ou LSA (latent semantic analysis) [Conover *et al.*, 2011a] au corpus afin d'extraire des traits thématiques.

Étant donnée l'hétérogénéité des performances obtenues par les différents travaux cherchant à classer les points de vue politiques sur Twitter, Cohen et Ruths [2013] ont étudié la difficulté de cette tâche en fonction de plusieurs types d'utilisateurs : les personalités politiques, les utilisateurs avec une activité politique soutenue sur Twitter (par exemple, adhésion à des listes politiques ou déclaration de l'orientation politique sur le profil) et les utilisateurs avec une activité politique modérée. Il en résulte ainsi que la tâche de classification de points de vue politique sur Twitter reste difficile pour les utilisateurs avec une activité politique modérée.

Il est également pertinent de signaler une autre limite des approches présentées dans cette section : leur dépendance vis-à-vis de données annotées. Par conséquent, d'autres travaux,

12. Les travaux mentionnés dans cette section se focalisent sur les points de vue politiques exprimés sur Twitter. Ceux-ci représentent, à notre connaissance, la majorité des travaux sur la classification de points de vue au niveau de l'utilisateur.

présentés dans la Section 2.3.2.2 ont abordé le problème d'identification du point de vue des utilisateurs de manière semi-supervisée ou non supervisée.

2.3.2.2 Approches semi-supervisées et non supervisées

À notre connaissance, la première approche non supervisée proposée dans la littérature pour regrouper les utilisateurs de médias sociaux en fonction de leur point de vue est celle de Abu-Jbara *et al.* [2012]. La méthode employée, similaire aux techniques de fouille d'opinions individuelles, comprend les étapes suivantes : identification des expressions d'opinions (formulées par les utilisateurs), de leur polarité et des cibles de ces opinions. Enfin, ces différents indices sont utilisés pour représenter chaque utilisateur par un vecteur de traits et un algorithme de regroupement k-moyennes est appliqué pour obtenir les points de vue des utilisateurs.

Alternativement, certains travaux ont proposé des méthodes de factorisation de matrices (*matrix factorization*) semi-supervisées [Gao *et al.*, 2014; Qiu *et al.*, 2015] et non supervisées [Gottipati *et al.*, 2014; Qiu *et al.*, 2013a] pour traiter cette même tâche. Ces méthodes présentent l'avantage de rendre possible la combinaison de différents facteurs tels que les interactions entre utilisateurs et les opinions qu'ils expriment dans le contenu posté pour identifier leur point de vue. Les approches décrites dans [Gottipati *et al.*, 2014; Qiu *et al.*, 2015, 2013a], définies spécifiquement pour les forums, définissent la nature des interactions entre utilisateurs à partir de la polarité du *post* qu'un utilisateur adresse en réponse au *post* d'un autre utilisateur. Dans [Gottipati *et al.*, 2014; Qiu *et al.*, 2013a], le contenu textuel des *posts* est traité au préalable afin d'en extraire la cible de l'opinion (par exemple, une entité nommée) et la polarité associée. Ainsi, ces modèles n'utilisent pas directement le texte mais plutôt les traits qui en ont été extraits. À l'inverse, les modèles proposés dans [Gao *et al.*, 2014; Qiu *et al.*, 2015] modélisent conjointement le phénomène d'occurrence des mots dans les *posts* et les interactions entre utilisateurs. En outre, l'approche semi-supervisée de factorisation de matrice développée par Qiu *et al.* [2015] a adopté une méthode comparable aux modèles thématiques en intégrant la notion de thème dans la découverte de points de vue.

Des modèles thématiques non supervisés inspirés par LDA, basés sur des distributions de Dirichlet et des variables discrètes, ont également été proposés pour identifier le point de vue des utilisateurs [Joshi *et al.*, 2016; Qiu et Jiang, 2013]. Les modèles JVTM (*joint viewpoint topic model*) introduits par Qiu et Jiang [2013] ont la particularité d'ajouter une dépendance du thème vis-à-vis du point de vue. En effet, une étude préliminaire présentée dans cet article a observé que des points de vue différents se focalisent sur des thèmes différents. Par exemple, les opposants au mariage homosexuel sont généralement plus enclins à mentionner la religion que ses partisans ne le sont. Les différentes versions de JVTM étant destinées à être appliquées à des forums organisés en fils de discussion, les interactions entre utilisateurs sont basées sur la polarité des réponses d'un utilisateur à un autre, de la même manière que [Gottipati *et al.*, 2014; Qiu *et al.*, 2015, 2013a]. Le modèle de Joshi *et al.* [2016] a quant à lui été développé pour Twitter. Il n'utilise pas les interactions entre utilisateurs mais exploite des indicateurs

d'opinion tels que les *emoticons*. Le modèle s'appuie également sur les parties de discours afin de faire la distinction entre mots d'opinion et mots thématiques.

Similairement à la méthode NOMINATE [Poole et Rosenthal, 1985] introduite dans la Section 2.3.1.2, d'autres travaux ont modélisé de manière non supervisée le point de vue des utilisateurs par des variables continues plutôt que discrètes [Bamman et Smith, 2015; Barberá, 2015]. Bamman et Smith [2015] ont proposé un modèle additif (*additive model*) capable de découvrir les points de vue latents des utilisateurs à partir d'observations sous la forme de couples (*sujet*, *prédicat*) (par exemple, (*Global warming*, *is a hoax*)). Ces couples sont extraits au préalable de manière non supervisée par un système d'extraction d'information. Ce modèle ne permet toutefois pas l'intégration d'éventuelles interactions entre utilisateurs. À l'inverse, le modèle décrit par Barberá [2015] se base exclusivement sur les interactions entre utilisateurs. Ce modèle est une adaptation aux utilisateurs de Twitter des *ideal point models*, normalement utilisés pour positionner les législateurs dans le spectre politique à partir de leur vote sur un ensemble de textes de loi. Il remplace ainsi la décision de vote des législateurs par le choix des utilisateurs de *follow* ou non un ensemble de personnalités politiques sur Twitter.

Le regroupement d'utilisateurs de médias sociaux à partir de leurs interactions et du contenu qu'ils ont généré a également été étudié sous un angle légèrement différent dans la tâche connue sous le nom de « détection de communautés » (*community detection*). Généralement associée à la théorie des graphes et des réseaux, la détection de communautés a pour objectif de découvrir des groupes de nœuds fortement connectés [Leskovec *et al.*, 2010]. Appliquée aux réseaux sociaux, la détection de communautés cherche à regrouper les utilisateurs qui interagissent beaucoup entre eux, par exemple les utilisateurs qui partagent un centre d'intérêt et échangent sur ce sujet. Ainsi, les techniques développées en détection de communautés peuvent également être utiles pour regrouper les utilisateurs en fonction de leur point de vue. En plus des méthodes basées uniquement sur la structure du réseau mentionnées dans [Leskovec *et al.*, 2010], des modèles thématiques ont également été proposés pour cette tâche afin d'exploiter conjointement les interactions et le contenu textuel posté par les utilisateurs [Liu *et al.*, 2016, 2014; Sachan *et al.*, 2012, 2014]. Les modèles introduits par Sachan *et al.* [2012, 2014] étendent LDA en ajoutant simplement des variables latentes correspondant à la communauté des utilisateurs et des variables observées indiquant si un utilisateur a interagi avec un autre utilisateur. Un modèle similaire est décrit dans [Liu *et al.*, 2016, 2014], à la différence près que ce dernier est doublement non-paramétrique : plutôt que fixer *a priori* le nombre de thèmes et de communautés, ces quantités sont apprises à partir des données pendant l'inférence.

Les différents travaux que nous avons décrits dans cette section se focalisent sur l'identification des points de vue exprimés dans les documents longs ou par les utilisateurs de médias sociaux. Alternativement, les points de vue peuvent être étudiés à un niveau plus global, que nous nommons « macroscopique » en considérant non plus les documents ou utilisateurs individuellement mais plutôt l'agrégat de documents ou d'utilisateurs. Cela mène à l'introduction de tâches différentes dont le but n'est plus d'identifier des groupes de point de vue mais plutôt d'analyser le discours des populations associées à ces groupes ou de détecter les sujets controversés (c'est-à-dire les sujets susceptibles de générer des groupes de point de vue). Nous

passons en revue la littérature portant sur la fouille de points de vue au niveau macroscopique dans la Section 2.4.

2.4 Fouille au niveau macroscopique : populations et sujets

À la différence des travaux au niveau mésoscopique décrits dans la Section 2.3, la fouille de points de vue au niveau macroscopique ne cherche pas assigner chaque document ou utilisateur à un point de vue. On considère au contraire les deux scénarios suivants et leurs travaux associés :

— l'analyse comparative de points de vue entre différentes populations, dans laquelle on considère que les attributions de points de vue sont déjà connues et où l'on étudie le discours propre à ces points de vue [Brigadir *et al.*, 2015; Chen *et al.*, 2015; Fang *et al.*, 2012; Lin *et al.*, 2008; Paul et Girju, 2009] (Figure 2.5a) ;

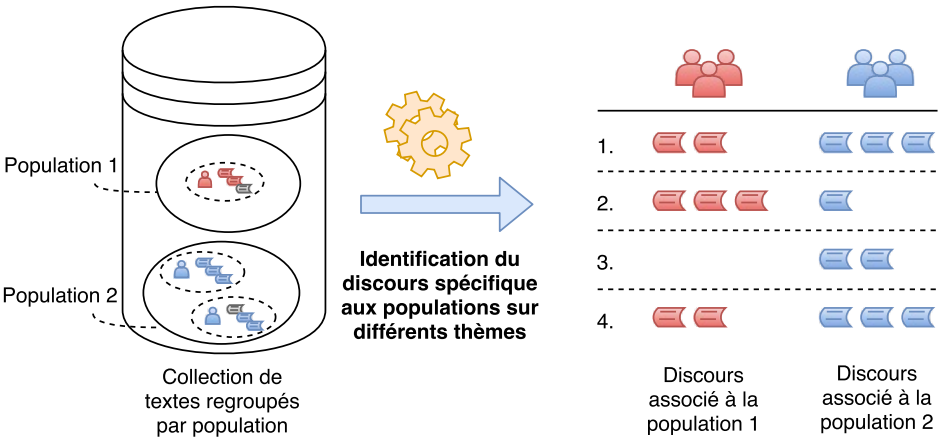
— la détection de sujets de controverse, qui n'a pas pour but d'identifier quel document ou utilisateur exprime quel point de vue, mais cherche plutôt à découvrir quels sujets sont controversés et quels sujets ne le sont pas [Balasubramanyan *et al.*, 2012; Dori-Hacohen et Allan, 2013, 2015; Garimella *et al.*, 2016; Guerra *et al.*, 2013; Jang et Allan, 2016; Jang *et al.*, 2016; Lewenberg *et al.*, 2016; Lin et Hauptmann, 2006; Popescu et Pennacchiotti, 2010; Wang *et al.*, 2014] (Figure 2.5b).

Nous détaillons ces différents travaux dans les Sections 2.4.1 et 2.4.2, respectivement.

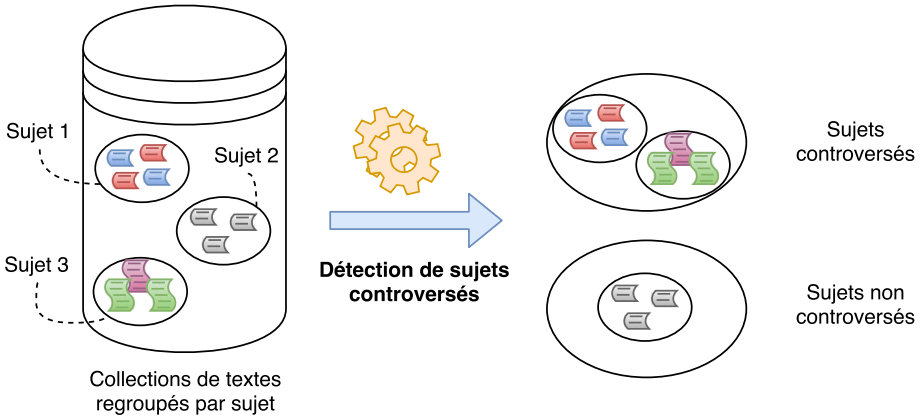
2.4.1 Analyse comparative des points de vue entre différentes populations

Comme nous l'avions illustré à travers l'exemple de la Section 2.1.3 sur la nuance entre « occupé les territoires palestiniens » et « aménagé des implantations », un même fait ou thème peut être discuté avec des mots distincts suivant le point de vue de l'auteur. Dans cette section, nous présentons ainsi les travaux qui décrivent des méthodes permettant de comparer le discours spécifique aux différents points de vue. Notons également que dans ces travaux le point de vue exprimé dans les documents ou par les utilisateurs est une information considérée comme connue *a priori* et fournie en entrée des algorithmes proposés.

À notre connaissance, Lin *et al.* [2008] ont été les premiers à développer une approche pour identifier le discours propre à différents points de vue. Leur approche est basée sur un modèle thématique dans lequel chaque mot d'un document est tiré suivant une distribution combinant un poids thématique et un poids idéologique par une fonction *softmax*. Ces poids sont distribués suivant des lois normales, ce qui distingue cette approche de la méthode classique de LDA reposant sur des lois de Dirichlet. Les points de vue comparés dans [Lin *et al.*, 2008] sont politiques et, en particulier, associés au conflit israélo-palestinien en se basant sur la collection Bitterlemons introduite par Lin *et al.* [2006]. Le discours associé aux points de vue politiques a également été étudié dans [Brigadir *et al.*, 2015] par des méthodes de sémantique distributionnelle (*distributional semantics*).



(a) Analyse comparative des points de vue entre différentes populations. L'objectif est d'identifier le discours caractéristique de différentes populations et comparer celui-ci sur un ensemble de thèmes, à partir des textes associés à chaque population.



(b) Détection de sujets de controverse. L'objectif est de distinguer les sujets controversés des sujets non controversés, chaque sujet étant représenté par une collection de textes.

FIGURE 2.5 – Tâches de fouille de points de vue au niveau macroscopique.

D'autres travaux ont également adopté des techniques basées sur les modèles thématiques mais pour aborder la notion de point de vue sous un angle légèrement différent : celui de la différence inhérente à la culture [Chen *et al.*, 2015; Fang *et al.*, 2012; Paul et Girju, 2009]. En effet, deux populations de cultures différentes sont susceptibles de percevoir un fait ou un événement de manières distinctes. Par exemple, une actualité concernant le Dalaï Lama ne connaîtra pas le même traitement dans la presse occidentale et la presse chinoise (par exemple, dans le New York Times [13] et Xinhua News [14], tout deux en anglais). Le modèle décrit par Paul et Girju [2009] est une simple extension de LDA dans laquelle les documents sont supposés provenir de différentes collections (chaque collection correspondant à une culture ou perspective différente). Chaque mot est considéré comme issu d'une distribution spécifique au thème uniquement (comme dans LDA) ou bien spécifique au thème et à la collection du

13. `https://www.nytimes.com/`
14. `http://www.xinhuanet.com/english/`

document dont il provient, en utilisant un mécanisme de variables de commutation latentes pour prendre cette décision automatiquement. À l'inverse, le modèle CPTM (*cross-perspective topic model*) proposé dans [Fang *et al.*, 2012] se base sur les parties du discours pour identifier les mots thématiques et les mots d'opinion (spécifiques à un point de vue et à un thème), à la manière des travaux en fouille d'opinions individuelles. Les noms sont ainsi considérés comme des mots thématiques et les adjectifs, adverbes et verbes sont considérés comme des mots d'opinions. Chen *et al.* [2015] ont pour leur part développé un modèle similaire à celui de Paul et Girju [2009] à la différence près qu'il remplace les distributions de Dirichlet par des processus hiérarchiques de Pitman-Yor, permettant d'intégrer le phénomène de loi de puissance (*power law*) observé dans le langage.

Les travaux mentionnés dans cette section permettent de comparer les variations dans le discours adopté par différents points de vue exprimés dans des corpus de textes sur des sujets définis. Ils font ainsi l'hypothèse que ces sujets sont susceptibles de soulever et mettre en opposition plusieurs points de vue, ou, en d'autres termes, que ces sujets sont controversés. Cependant, tout sujet n'est pas nécessairement controversé. La détection de sujets de controverse est ainsi un problème à part entière que nous étudions dans la Section 2.4.2.

2.4.2 Détection de sujets de controverse

Un sujet de controverse peut être défini comme un sujet « provoquant une discussion publique dans laquelle les membres de l'audience expriment des opinions opposées, de la surprise ou de l'incrédulité » (traduit de l'anglais depuis [Popescu et Pennacchiotti, 2010]). Par exemple, des sujets tels que l'avortement, le mariage homosexuel et la vaccination sont caractérisés par le clivage existant entre les protagonistes qui s'expriment sur chacun de ces sujets. La détection de sujets de controverse peut être alors vue comme un moyen de lutter contre les phénomènes de chambres d'échos [Sunstein, 2009] et de bulles de filtre [Pariser, 2011]. En effet, l'identification des sujets de controverse pourrait permettre d'avertir les internautes en leur signalant que le sujet qu'ils consultent (par exemple, par l'intermédiaire d'un moteur de recherche) est susceptible de soulever des points de vue opposés et donc que les documents s'y rapportant devraient être lus avec précaution.

Dans la littérature, la détection de controverses a été abordée à la fois par des méthodes supervisées [Balasubramanyan *et al.*, 2012; Popescu et Pennacchiotti, 2010; Wang *et al.*, 2014] et par des méthodes non supervisées [Dori-Hacohen et Allan, 2013, 2015; Garimella *et al.*, 2016; Guerra *et al.*, 2013; Jang et Allan, 2016; Jang *et al.*, 2016]. Pour détecter sur Twitter les événements controversés (par exemple, l'accusation de viol portée sur David Copperfield entre 2007 et 2010 [15]), Popescu et Pennacchiotti [2010] ont proposé un classifieur de type arbre de décision et un ensemble de traits tels que les parties de discours, la présence de mots issus de lexiques d'opinion ou de controverse et les interactions entre utilisateurs (*retweet* et *reply*). Balasubramanyan *et al.* [2012] ont quant à eux étendu le modèle LDA supervisé pour prédire le niveau de controverse associé à différents thèmes, discutés dans des commentaires de blogs

15. `http://www.rtl.be/info/monde/france/david-copperfield-implique-dans-une-affaire-de-viol--21473.aspx`

politiques. Des classifieurs de type machine à vecteur de support et régression logistique ont également été proposés dans [Wang *et al.*, 2014] pour détecter les disputes dans les discussions de pages Wikipédia [16].

D'autres travaux ont également exploité Wikipédia pour détecter de manière non supervisée les pages Web traitant d'un sujet controversé ou non [Dori-Hacohen et Allan, 2013, 2015; Jang et Allan, 2016; Jang *et al.*, 2016]. Dori-Hacohen et Allan [2013, 2015]; Jang et Allan [2016] ont proposé d'aligner les pages Web aux pages Wikipédia en postulant qu'une page traite d'un sujet controversé si la page Wikipédia décrivant ce sujet est elle-même controversée. La nature controversée ou non d'une page Wikipédia est quant à elle détectée automatiquement en se basant sur les métadonnées et discussions associées à la page. Dans [Jang *et al.*, 2016], les auteurs ont construit un modèle de langue des sujets controversés appris sur les articles Wikipédia et utilisé ensuite pour identifier si une page Web est controversée.

La détection de controverses dans les réseaux sociaux a également été abordée sans supervision en se basant les interactions entre les différents utilisateurs [Garimella *et al.*, 2016; Guerra *et al.*, 2013]. Ces deux travaux ont défini des mesures basées sur la topologie du réseau pour quantifier le niveau de controverse d'un sujet. La méthode adoptée par Guerra *et al.* [2013] se base sur la mesure de modularité originalement proposée par Newman [2006]. Garimella *et al.* [2016] ont pour leur part proposé des approches de mesure alternatives également basées sur le réseau, telles que la marche aléatoire (*random walk*), le plongement de graphe (*graph embedding*) et le moment dipolaire (*dipole moment*). Les auteurs ont par ailleurs testé des méthodes simples basées sur le contenu et ont noté leur inefficacité par rapport aux méthodes basées sur le graphe d'utilisateurs.

2.5 Conclusion

Au cours de ce chapitre, nous avons passé en revue les différents travaux de l'état de l'art traitant de la notion de point de vue. Nous avons découpé ces travaux suivant trois niveaux : le niveau microscopique, le niveau mésoscopique et le niveau macroscopique. Au niveau microscopique, nous avons décrit les approches permettant de détecter les expressions d'argumentation ou de contention ainsi que celles employées pour classer les documents courts selon leur point de vue. Les travaux positionnés au niveau mésoscopique ont quant à eux abordé l'identification de points de vue dans les documents longs tels que les essais et les textes législatifs, ou exprimés par des utilisateurs dans les médias sociaux. Enfin, les tâches de fouille de points de vue au niveau macroscopique se focalisent sur l'analyse comparative du discours adopté par différentes populations (chacune associée à un point de vue) et la détection de sujets de controverse.

Parmi les méthodes employées pour résoudre ces différents problèmes, on distingue les méthodes supervisées, généralement basées sur des classifieurs tels que les machines à vecteurs de support et les réseaux de neurones, et les méthodes non supervisées, majoritai-

16. `https://www.wikipedia.org/`

rement construites sur des modèles thématiques. Les modèles thématiques constituent une approche polyvalente pour modéliser et extraire sans annotations préalables des thèmes ainsi que d'autres dimensions latentes telles que les points de vue à partir d'un corpus de textes. L'annotation d'un corpus étant souvent coûteuse et difficile à obtenir, nous avons décidé de baser nos approches de fouille de points de vue sur les modèles thématiques et en particulier ceux inspirés de LDA, que nous décrivons en détail dans le Chapitre 3.

Chapitre 3

Modèles thématiques probabilistes

Sommaire

3.1 Introduction

Les modèles thématiques (*topic models*) regroupent l'ensemble des approches dont le but est de découvrir de manière non supervisée les thèmes latents apparaissant dans une collection de documents textuels. Cette tâche est parfois connue sous le nom d'indexation sémantique latente (LSI – *latent semantic indexing*) ou d'analyse sémantique latente (LSA – *latent semantic analysis*), provenant du premier modèle thématique proposé dans [Deerwester *et al.*, 1990]. L'approche fondatrice de Deerwester *et al.* [1990] consiste à former une matrice d'occurrences terme-document pour la collection concernée, puis à lui appliquer une décomposition en valeurs singulières (SVD – *singular value decomposition*). Enfin, en ne conservant que les T plus grandes valeurs singulières, on obtient une approximation de rang inférieure de la matrice d'occurrences. Le nombre T désigne alors le nombre de thèmes et est choisi très petit devant la taille du vocabulaire. Toutefois, le modèle LSI représente les documents et les termes par des vecteurs à valeurs réelles (pouvant être positives ou négatives) dans l'espace des thèmes de dimension T. Il est alors difficile d'interpréter les composantes négatives de ces vecteurs. Pour pallier ce problème, Hofmann a proposé une version probabiliste de LSI, intitulée pLSI (*probabilistic latent semantic indexing*), qui présente l'avantage de n'utiliser

que des vecteurs à valeurs positives. Cependant, pLSI n'est pas un modèle génératif – il ne permet pas de calculer naturellement la probabilité d'un nouveau document – et est sujet au surapprentissage.

Pour surmonter ces différentes limites, Blei *et al.* [2001, 2003] ont proposé l'allocation de Dirichlet latente (LDA – *latent Dirichlet allocation*), la version générative et bayésienne de pLSI. En raison de sa flexibilité, LDA a par la suite connu une grande popularité et a été adapté à diverses applications [Boyd-Graber *et al.*, 2017]. Par exemple, des extensions de LDA ont été proposées pour modéliser conjointement les thèmes et les opinions [Jo et Oh, 2011; Lin et He, 2009], pour décrire la dynamique temporelle des thèmes [Blei et Lafferty, 2006; Wang et McCallum, 2006] ou pour découvrir les communautés dans les réseaux sociaux [Sachan *et al.*, 2012]. Nous avons également développé des extensions de LDA pour modéliser les points de vue, qui seront détaillés dans les Chapitres 4 et 5. Le chapitre présent fait ainsi office d'introduction au modèle LDA et plus généralement aux aspects et méthodes propres à ses extensions. Dans la Section 3.2, nous fournissons une description formelle du modèle LDA. La Section 3.3 aborde ensuite la question de l'inférence du modèle LDA et de ses extensions. Enfin, la Section 3.4 présente les différentes méthodes utilisées dans la littérature pour évaluer et comparer les performances des modèles thématiques [1].

3.2 LDA : allocation de Dirichlet latente

L'allocation de Dirichlet latente (LDA) [Blei *et al.*, 2001, 2003] est un modèle probabiliste bayésien et génératif capable de découvrir de manière non supervisée les différents thèmes latents apparaissant dans une collection de documents textuels. Cette section décrit en détail le modèle LDA, qui constitue la base sur laquelle nos contributions s'appuient. Les notations adoptées dans cette section sont résumées dans le Tableau 3.1 et seront également introduites lors de leur première utilisation.

3.2.1 Histoire générative

LDA, en tant que modèle génératif, peut être défini formellement en déroulant le processus de génération d'une collection de documents donnée selon ce modèle. La description de ce processus est connu sous le nom d'histoire générative (*generative story*) dans la littérature. Présentons tout d'abord de manière informelle l'histoire générative de LDA. Selon LDA, chaque document $d = 1, \ldots, D$ a une propension plus ou moins grande à traiter un ensemble de thèmes $z = 1, \ldots, T$. Cette propension de d pour le thème z est dénotée par $\theta_{dz} \in [0, 1]$. De manière similaire, chaque mot $w = 1, \ldots, W$ du vocabulaire apparaîtra plus ou moins probablement en fonction des thèmes $z = 1, \ldots, T$ abordés. La probabilité de voir le mot w apparaître pour le thème z est mesurée par la quantité $\phi_{zw} \in [0, 1]$. Par ailleurs, LDA adopte une posture bayésienne en traitant les distributions $\{\boldsymbol{\theta}_d\}_{d=1}^{D}$ et $\{\boldsymbol{\phi}_z\}_{z=1}^{T}$ comme des variables

1. Dans le reste de ce chapitre, nous désignerons les modèles thématiques probabilistes simplement comme « modèles thématiques » par soucis de concision.

TABLEAU 3.1 – Notations adoptées pour le modèle LDA.

Symbole	Définition
D	Nombre de documents dans la collection.
N_d	Nombre de mots dans le document d.
W	Taille du vocabulaire.
T	Nombre de thèmes.
w_{dn}	Le n^{e} mot du document d.
z_{dn}	Le thème attribué au n^{e} mot du document d.
$\mathbf{w}$	Vecteur de tous les mots de la collection.
$\mathbf{z}$	Vecteur de toutes les attributions de thèmes.
$\boldsymbol{\theta}$	Matrice de taille $D \times T$ contenant les distributions de thèmes spécifiques aux documents.
$\boldsymbol{\phi}$	Matrice de taille $T \times W$ contenant les distributions de mots spécifiques aux thèmes.
α	Paramètre du *prior* de Dirichlet symétrique sur $\boldsymbol{\theta}$.
β	Paramètre du *prior* de Dirichlet symétrique sur $\boldsymbol{\phi}$.
n_{dz}	Nombre de mots dans le document d attribués au thème z.
n_{zw}	Nombre d'occurrences du mot w attribuées au thème z.

aléatoires, tirées de lois de Dirichlet symétriques respectivement de dimension T (nombre de thèmes) et W (taille du vocabulaire) et de paramètre α et β.

On suppose maintenant que chaque document d contient N_d mots qui doivent être générés. Le document d est alors construit mot par mot. Pour chaque emplacement de mot $n = 1, ..., N_d$, le thème z_{dn} (qui va être attribué au n^{e} mot de d) est d'abord choisi aléatoirement suivant la loi discrète [2] paramétrisée par le vecteur $\boldsymbol{\theta}_d = (\theta_{dz})_{z=1}^{T}$. Ensuite, le mot w_{dn} est tiré suivant la loi discrète paramétrisée par le vecteur $\boldsymbol{\phi}_{z_{dn}} = (\phi_{z_{dn}w})_{w=1}^{W}$, dépendant de l'attribution de thème z_{dn}. Intuitivement, cela traduit le fait que l'auteur d'un document détermine d'abord le thème associé à chaque mot du document, puis « verbalise » ce thème en choisissant un mot du vocabulaire spécifique à ce thème.

Plus formellement, l'histoire générative de LDA peut être décrite de la manière suivante :

1. Tirer une distribution de mots $\boldsymbol{\phi}_z \sim \text{Dirichlet}_W(\beta)$ pour chaque thème $z = 1, ..., T$;
2. Pour chaque document $d = 1, ..., D$:
   (a) Tirer une distribution de thèmes $\boldsymbol{\theta}_d \sim \text{Dirichlet}_T(\alpha)$;
   (b) Pour chaque emplacement de mots $n = 1, ..., N_d$:
      i. Tirer un thème $z_{dn} \sim \text{Discrète}(\boldsymbol{\theta}_d)$;
      ii. Tirer un mot $w_{dn} \sim \text{Discrète}(\boldsymbol{\phi}_{z_{dn}})$.

2. Dans la littérature, cette loi est parfois également nommée loi catégorielle ou, par abus de langage, loi multinomiale.

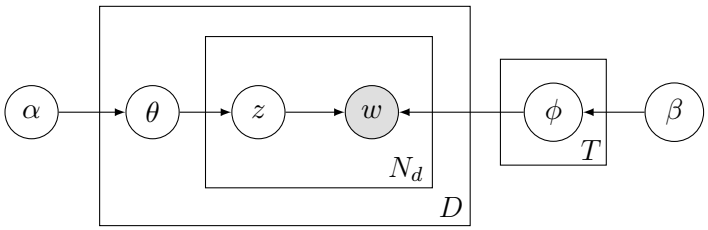


FIGURE 3.1 – Représentation de LDA sous forme de modèle graphique.

3.2.2 Représentation sous forme de modèle graphique

L'histoire générative de LDA peut également être représentée sous forme de modèle graphique, comme illustré dans la Figure 3.1. Dans ce type de schéma, les cercles décrivent les variables aléatoires – grisées pour les variables observées et blanches pour les variables latentes. Les rectangles, appelées *plates* dans la littérature, dénotent des réplications de variables aléatoires avec pour cardinalité le nombre indiqué dans le coin inférieur droit [3]. Par exemple, la variable w est répliquée $\sum_{d=1}^{D} N_d$ fois et désigne donc $\left\{\{w_{dn}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$, c'est-à-dire tous les mots des documents de la collection.

Une flèche entre deux cercles indique une dépendance statistique entre les variables correspondantes. Par exemple, la variable w_{dn} dépend de ses parents z_{dn} et $\{\boldsymbol{\phi}_z\}_{z=1}^{T}$ (ou, plus précisément, de $\boldsymbol{\phi}_{z_{dn}}$). Cela permet également de déduire les relations d'indépendance conditionnelle entre les différentes variables : conditionnellement à z_{dn}, $\{\boldsymbol{\phi}_z\}_{z=1}^{T}$ et $\mathbf{w}^{-(dn)}$ (l'ensemble des mots de la collection à l'exception de w_{dn}), w_{dn} est indépendant de toutes les autres variables aléatoires. On dit alors que $\{z_{dn}, \{\boldsymbol{\phi}_z\}_{z=1}^{T}, \mathbf{w}^{-(dn)}\}$ est la couverture de Markov (*Markov blanket*) de la variable w_{dn}.

Notons toutefois que la représentation d'un modèle thématique sous forme de modèle graphique, bien qu'utile pour donner au lecteur l'intuition immédiate du modèle, ne délivre pas autant d'informations que l'histoire générative. En effet, les lois de probabilités (par exemple, lois de Dirichlet et lois discrètes dans le cas de LDA) sont généralement omises dans le modèle graphique. De plus, le modèle graphique manque de précision pour décrire la nature d'une dépendance liant deux variables aléatoires : z et ϕ sont seulement représentés comme deux parents de w, sans expliciter la formule du tirage de w comme le fait l'histoire générative. La représentation sous forme de modèle graphique ne doit donc pas être perçue comme un moyen de définir un modèle thématique – l'histoire générative demeure pour cela indispensable – mais plutôt comme un moyen de l'illustrer.

3.2.3 Vraisemblance et probabilité postérieure du modèle

Afin de rendre LDA applicable, il est nécessaire de calculer la probabilité postérieure du modèle, c'est-à-dire la probabilité des variables latentes conditionnellement aux variables

3. Cette notation est couramment adoptée dans la littérature pour fournir une représentation plus compacte et lisible des modèles.

observées $p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{w}\,;\alpha,\beta)$. En effet, connaître cette distribution permettrait de découvrir concrètement les thèmes les plus probablement assignés aux différents documents et les mots les plus probablement attribués aux différents thèmes. Cette probabilité postérieure peut être décomposée de la manière suivante :

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}|\mathbf{w}\,;\alpha,\beta) = \frac{p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}\,;\alpha,\beta)}{p(\mathbf{w}\,;\alpha,\beta)}. \tag{3.1}$$

Le numérateur est la probabilité jointe de toutes les variables aléatoires (latentes et observées) et le dénominateur correspond à la vraisemblance du modèle.

Le calcul de la probabilité jointe $p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}\,;\alpha,\beta)$ résulte de la simple utilisation des dépendances illustrées par le modèle graphique de la Figure 3.1 et du remplacement des lois de probabilité par leur définition :

$$\begin{aligned} & p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}\,;\alpha,\beta) \\ &= \left\{\prod_{d=1}^{D}\prod_{n=1}^{N_d} p(w_{dn}|\boldsymbol{\phi}_{z_{dn}})\, p(z_{dn}|\boldsymbol{\theta}_d)\right\}\left\{\prod_{d=1}^{D} p(\boldsymbol{\theta}_d\,;\alpha)\right\}\left\{\prod_{z=1}^{T} p(\boldsymbol{\phi}_z\,;\beta)\right\} \\ &= \left\{\prod_{d=1}^{D}\prod_{n=1}^{N_d} \phi_{z_{dn}w_{dn}}\,\theta_{dz_{dn}}\right\}\left\{\prod_{d=1}^{D} \mathrm{Dirichlet}_T(\boldsymbol{\theta}_d\,;\alpha)\right\}\left\{\prod_{z=1}^{T} \mathrm{Dirichlet}_W(\boldsymbol{\phi}_z\,;\beta)\right\}. \end{aligned} \tag{3.2}$$

La vraisemblance du modèle $p(\mathbf{w}\,;\alpha,\beta)$ pose cependant problème. Son calcul nécessite la marginalisation (c'est-à-dire la sommation ou l'intégration) des variables latentes $\mathbf{z}$, $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$, ce qui mène à une intégrale qui ne peut être calculée en raison du couplage des variables $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$ causé par la marginalisation des attributions de thèmes $\mathbf{z}$ [4] :

$$\begin{aligned} & p(\mathbf{w}\,;\alpha,\beta) \\ &= \int_{\boldsymbol{\theta}}\int_{\boldsymbol{\phi}} \sum_{\mathbf{z}} p(\mathbf{w}, \mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}\,;\alpha,\beta)\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\phi} \\ &= \int_{\boldsymbol{\theta}}\int_{\boldsymbol{\phi}} \left\{\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{z_{dn}=1}^{T} \phi_{z_{dn}w_{dn}}\,\theta_{dz_{dn}}\right\}\left\{\prod_{d=1}^{D} p(\boldsymbol{\theta}_d\,;\alpha)\right\}\left\{\prod_{z=1}^{T} p(\boldsymbol{\phi}_z\,;\beta)\right\}\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{\phi}. \end{aligned} \tag{3.3}$$

Par conséquent, l'inférence postérieure du modèle LDA requiert l'utilisation de méthodes approchées, que nous décrivons dans la Section 3.3.

3.3 Méthodes d'inférence postérieure approchées

Dans cette section, nous présentons les principales méthodes approchées utilisées pour réaliser l'inférence de modèles thématiques bayésiens tels que LDA et ses extensions. La

4. Marginaliser $\mathbf{z}$ consiste à sommer sur l'ensemble des attributions possibles de chacune des variables z_{dn} pour $d = 1, ..., D$ et $n = 1, ..., N_d$. Pour simplifier l'écriture de cette somme complexe, nous introduisons la notation $\sum_{\mathbf{z}}$.

Section 3.3.1 détaille l'échantillonnage de Gibbs, méthode d'inférence la plus populaire [5] et que nous avons appliquée aux nouveaux modèles que nous proposons dans les Chapitres 4 et 5. Dans la Section 3.3.2, nous abordons brièvement la méthode variationnelle, qui constitue une approche alternative pour l'inférence de modèles thématiques.

3.3.1 Échantillonnage de Gibbs

La méthode d'inférence postérieure nommée « échantillonnage de Gibbs » (*Gibbs sampling*) fait partie des méthodes de Monte Carlo par chaînes de Markov (*Markov Chain Monte Carlo* – MCMC). De manière générale, les méthodes MCMC ont pour objectif de faciliter le tirage de variables aléatoires distribuées suivant des lois de probabilité multivariées complexes ou de calculer des quantités (par exemple, des intégrales) contenant de telles variables, qui ne sont pas calculables de manière analytique. Dans cette section, nous décrivons en particulier l'échantillonnage de Gibbs qui est souvent adopté pour l'inférence des modèles thématiques en raison de sa simplicité d'utilisation et de son insensibilité aux optima locaux (contrairement aux méthodes variationnelles).

3.3.1.1 Principe général

Décrivons tout d'abord l'échantillonnage de Gibbs de manière générique. Soit un ensemble de variables aléatoires $\mathbf{x} = \{x_i\}_{i=1}^N$. On considère comme connue (ou calculable analytiquement à une variable multiplicative près) la loi conditionnelle $p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)$, pour chaque $i = 1, ..., N$. On souhaite calculer la loi jointe $p(\mathbf{x})$, qui est une loi multivariée potentiellement de grande dimension et pour laquelle on ne dispose pas d'expression analytique. L'échantillonneur de Gibbs va construire une chaîne de Markov dont la distribution stationnaire est $p(x_1, ..., x_N)$, permettant ainsi d'effectuer des tirages approximativement distribués selon la loi jointe.

L'échantillonneur de Gibbs dans ce cas générique est illustré dans l'Algorithme 1. On suppose disposer des lois conditionnelles $p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)$ et d'un nombre de tirages souhaités de la loi jointe $p(\mathbf{x})$ (correspondant au nombre d'itérations). L'état initial de la chaîne de Markov correspond aux variables $x_1^{(0)}, ..., x_N^{(0)}$, initialisées aléatoirement (par exemple en utilisant une loi uniforme). Une boucle sur $s = 1, ..., S$ construit ensuite l'état s à partir de l'état précédent $s-1$. Chaque variable $x_i^{(s)}$ pour $i = 1, ..., N$ est successivement tirée à partir de sa loi conditionnelle $p(x_i|x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_N^{(s-1)})$. On note que les variables utilisées pour le calcul de la probabilité conditionnelle de $x_i^{(s)}$ sont les variables de l'état s pour les indices inférieurs à i (ces variables ont déjà été tirées pour l'état s) et celles de l'état $s-1$ pour les indices supérieurs à i (ces variables n'ont pas encore été tirées pour l'état

5. En analysant manuellement les articles proposant de nouveaux modèles thématiques et publiés dans les conférences CIKM, KDD, SIGIR, WSDM et WWW de 2013 à 2016, nous avons identifié un total de 110 articles parmi lesquels 73 ont utilisé l'échantillonnage de Gibbs et 11 se sont basés sur la méthode variationnelle.

Algorithme 1 : Échantillonneur de Gibbs générique

Input : un nombre d'itérations S, les lois conditionnelles
$p(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_N)$, pour chaque $i = 1, ..., N$
Output : S tirages approximativement distribués selon $p(x_1, ..., x_N)$
Initialiser $x_1^{(0)}, ..., x_N^{(0)}$ aléatoirement
for $s \leftarrow 1$ **to** S **do**
  for $i \leftarrow 1$ **to** N **do**
    Tirer $x_i^{(s)} \sim p(x_i|x_1^{(s)}, ..., x_{i-1}^{(s)}, x_{i+1}^{(s-1)}, ..., x_N^{(s-1)})$
  end
end
return $\left\{\mathbf{x}^{(s)}\right\}_{s=1}^{S}$

s). Finalement, on obtient un ensemble de S tirages (un tirage par état) approximativement distribués suivant la loi jointe $p(\mathbf{x})$.

Une des applications de cet échantillonneur de Gibbs est de rendre possible le calcul d'intégrales telles que la suivante :

$$I = \int_{\mathbf{x}} f(\mathbf{x})\, p(\mathbf{x})\, \mathrm{d}\mathbf{x} \tag{3.4}$$

où f est une fonction quelconque telle que $I < \infty$. La probabilité jointe $p(\mathbf{x})$ n'étant pas connue, il n'est pas possible de calculer I analytiquement. Cependant, les S tirages collectés par un échantillonneur de Gibbs permettent d'approcher I de la manière suivante :

$$I \approx \frac{1}{S} \sum_{s=1}^{S} f(\mathbf{x}^{(s)}). \tag{3.5}$$

Lors de l'utilisation d'un échantillonneur de Gibbs, il est cependant nécessaire de prendre en considération les points suivants :

— Les premiers tirages ne sont généralement pas distribués suivant la loi jointe car la chaîne de Markov nécessite un certain nombre d'itérations préliminaires pour converger. Pour cette raison, on ne collecte en pratique pas les premiers tirages. On laisse l'algorithme tourner pendant un certain nombre d'itérations B (période dite de *burn-in*) et, seulement après cette période, on collecte les tirages – qui, cette fois, sont plus vraisemblablement tirés de la loi jointe.

— Les tirages provenant d'états successifs sont corrélés. En effet, le tirage des variables $\left\{x_i^{(s)}\right\}_{i=1}^{N}$ s'effectue en conditionnant sur les variables $\left\{x_i^{(s-1)}\right\}_{i=1}^{N}$. Afin d'éviter d'obtenir des tirages corrélés, on ne collecte que les tirages d'états séparés par L itérations – les tirages d'états plus « distants » dans la chaîne de Markov ont moins de chance d'être corrélés.

En plus de la corrélation entre états successifs d'un échantillonneur de Gibbs, il est également possible d'avoir des corrélations entre les variables $x_1, ..., x_N$. Ce second type de corréla-

Algorithme 2 : Échantillonneur de Gibbs marginalisé

Input : un nombre d'itérations S, les lois conditionnelles
$p(z_k|z_1, ..., z_{k-1}, z_{k+1}, ..., z_P)$, pour chaque $k = 1, ..., P$
Output : S tirages approximativement distribués selon $p(z_1, ..., z_P)$
Initialiser $z_1^{(0)}, ..., z_P^{(0)}$ aléatoirement
for $s \leftarrow 1$ **to** S **do**
  for $k \leftarrow 1$ **to** P **do**
    Tirer $z_k^{(s)} \sim p(z_k|z_k^{(s)}, ..., z_{k-1}^{(s)}, z_{k+1}^{(s-1)}, ..., z_P^{(s-1)})$
  end
end
return $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$

tion a pour effet de réduire la capacité de « mélange » (*mixing*) de la chaîne de Markov, ce qui signifie que l'exploration de l'espace des variables $x_1, ..., x_N$ sera plus lente. Cela peut s'avérer problématique car l'utilisation de tirages ne représentant pas la loi cible mènerait à des estimations faussées. Il est alors nécessaire d'augmenter la valeur de S pour s'assurer de collecter des tirages parcourant la totalité de l'espace des variables $x_1, ..., x_N$, ce qui peut considérablement impacter le temps d'exécution. Afin de limiter certaines corrélations entre les variables, une version spéciale de l'échantillonneur de Gibbs a été proposée : l'échantillonneur de Gibbs marginalisé. Nous décrivons cette méthode dans la Section 3.3.1.2.

3.3.1.2 Échantillonnage de Gibbs marginalisé

L'échantillonnage de Gibbs marginalisé (*collapsed Gibbs sampling*) [Liu, 1994] consiste à marginaliser (c'est-à-dire intégrer) certaines variables afin d'effectuer des tirages uniquement pour les variables restantes. Reprenons l'exemple précédent où l'on cherche à estimer la probabilité jointe $p(x_1, ..., x_N)$. Supposons que les variables $\mathbf{x} = \{x_i\}_{i=1}^N$ peuvent être séparées en deux ensembles de variables $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^M$ et $\mathbf{z} = \{z_k\}_{k=1}^P$ tels que $N = M + P$ et tels que l'on peut calculer de manière analytique les probabilités conditionnelles suivantes :

$$p(z_k|z_1, ..., z_{k-1}, z_{k+1}, ..., z_P) = \int_{\boldsymbol{\theta}} p(z_k|\boldsymbol{\theta})\, p(\boldsymbol{\theta}|z_1, ..., z_{k-1}, z_{k+1}, ..., z_P)\, \mathrm{d}\boldsymbol{\theta} \tag{3.6}$$

pour chaque $k = 1, ..., P$. L'échantillonneur de Gibbs marginalisé résultant, résumé dans l'Algorithme 2, consiste alors simplement à tirer successivement les variables $\{z_k\}_{k=1}^P$ à partir de leur probabilité conditionnelle décrite dans l'Équation (3.6). Après collection des tirages $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$, la loi jointe de $\boldsymbol{\theta}$ peut être approchée de la manière suivante :

$$p(\boldsymbol{\theta}) \approx \frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{\theta}|\mathbf{z}^{(s)}). \tag{3.7}$$

L'échantillonneur de Gibbs marginalisé présente plusieurs avantages par rapport à la version générique :

— P étant inférieur à N, la complexité d'un échantillonneur de Gibbs marginalisé est moindre que celle d'un échantillonneur de Gibbs générique.

— La corrélation entre les variables $\{\theta_j\}_{j=1}^{M}$ et $\{z_k\}_{k=1}^{P}$ n'a plus d'impact sur la capacité de mélange de la chaîne de Markov, car l'échantillonneur de Gibbs marginalisé intègre les variables $\{\theta_j\}_{j=1}^{M}$ et effectue des tirages seulement pour les variables $\{z_k\}_{k=1}^{P}$.

Dans la Section 3.3.1.3, nous décrivons l'application de l'échantillonnage de Gibbs marginalisé à l'inférence postérieure du modèle LDA.

3.3.1.3 Application à LDA

Le modèle LDA contient les variables latentes suivantes : les attributions de thèmes à chaque mot $\mathbf{z} = \left\{\{z_{dn}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$, les distributions de thèmes spécifiques aux documents $\boldsymbol{\theta} = \{\boldsymbol{\theta}_d\}_{d=1}^{D}$ et les distributions de mots spécifiques aux thèmes $\boldsymbol{\phi} = \{\boldsymbol{\phi}_z\}_{z=1}^{T}$. Les seules variables observées sont les mots de la collection de documents $\mathbf{w} = \left\{\{w_{dn}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$. L'échantillonneur de Gibbs marginalisé de LDA [Griffiths et Steyvers, 2004] consiste à intégrer les distributions $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$ et à effectuer des tirages successifs de chaque variable z_{dn} pour $d = 1, ..., D$ et $n = 1, ..., N_d$ conditionnellement aux autres variables latentes $\mathbf{z}^{-(dn)}$ (c'est-à-dire toutes les attributions de thèmes sauf celle du mot courant) et aux observations $\mathbf{w}$. L'élément clé de l'échantillonneur de Gibbs marginalisé de LDA est donc le calcul de la probabilité $p(z_{dn}|\mathbf{z}^{-(dn)}, \mathbf{w}\,;\alpha,\beta)$, qui peut être écrite comme suit :

$$p(z_{dn} = z|\mathbf{z}^{-(dn)}, \mathbf{w}\,;\alpha,\beta) \propto \frac{p(\mathbf{w}, \mathbf{z}\,;\alpha,\beta)}{p(\mathbf{w}^{-(dn)}, \mathbf{z}^{-(dn)}\,;\alpha,\beta)} \tag{3.8}$$

Dans le reste de cette section, nous détaillons le développement aboutissant à la forme simplifiée de cette probabilité. La dérivation mathématique de l'échantillonneur de Gibbs marginalisé que nous présentons ici est inspirée de [Heinrich, 2008].

Calcul de $p(\mathbf{w}, \mathbf{z}\,;\alpha,\beta)$ La première étape du calcul de $p(z_{dn}|\mathbf{z}^{-(dn)}, \mathbf{w}\,;\alpha,\beta)$ consiste à déterminer la probabilité jointe de $\mathbf{w}$ et $\mathbf{z}$ après marginalisation de $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$:

$$p(\mathbf{w}, \mathbf{z}\,;\alpha,\beta) = p(\mathbf{w}|\mathbf{z}\,;\beta)\, p(\mathbf{z}\,;\alpha). \tag{3.9}$$

Le membre de droite est composé de deux termes : le premier terme est une vraisemblance et le second terme est une probabilité *a priori* (un *prior*). Le premier terme peut être calculé comme suit :

$$\begin{aligned} p(\mathbf{w}|\mathbf{z}\,;\beta) &= \int_{\boldsymbol{\phi}} p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi})\, p(\boldsymbol{\phi}\,;\beta)\, \mathrm{d}\boldsymbol{\phi} \\ &= \int_{\boldsymbol{\phi}} \left\{ \prod_{d=1}^{D} \prod_{n=1}^{N_d} \phi_{z_{dn} w_{dn}} \right\} \left\{ \prod_{z=1}^{T} \mathrm{Dirichlet}_W(\boldsymbol{\phi}_z\,;\beta) \right\} \mathrm{d}\boldsymbol{\phi} \end{aligned} \tag{3.10}$$

$$
\begin{aligned}
&= \int_{\boldsymbol{\phi}} \left\{ \prod_{z=1}^{T} \prod_{w=1}^{W} (\phi_{zw})^{n_{zw}} \right\} \left\{ \prod_{z=1}^{T} \frac{1}{\mathrm{B}(\beta)} \prod_{w=1}^{W} (\phi_{zw})^{\beta-1} \right\} \mathrm{d}\boldsymbol{\phi} \\
&= \prod_{z=1}^{T} \frac{\mathrm{B}(\mathbf{n}_z + \beta)}{\mathrm{B}(\beta)} \int_{\boldsymbol{\phi}_z} \frac{1}{\mathrm{B}(\mathbf{n}_z + \beta)} \prod_{w=1}^{W} (\phi_{zw})^{n_{zw}+\beta-1} \mathrm{d}\boldsymbol{\phi}_z \\
&= \prod_{z=1}^{T} \frac{\mathrm{B}(\mathbf{n}_z + \beta)}{\mathrm{B}(\beta)}
\end{aligned}
$$

où n_{zw} désigne le compteur du nombre de mots w attribués au thème z dans la collection, $\mathbf{n}_z$ est le vecteur $(n_{zw})_{w=1}^{W}$ et $\mathrm{B}(\cdot)$ désigne la fonction Beta d'Euler multivariée [6]. La deuxième ligne résulte de la simple application de la définition de $p(\mathbf{w}|\mathbf{z}, \boldsymbol{\phi})$ et $p(\boldsymbol{\phi}\,;\beta)$. Le passage de la deuxième ligne à la troisième est effectué en regroupant ensemble les mots de même type w et avec le même thème z attribué, en comptant les multiples occurrences avec n_{zw}. On observe alors dans la quatrième ligne qu'en regroupant tous les termes ϕ_{zw} et en introduisant $\mathrm{B}(\mathbf{n}_z + \beta)$, on obtient l'intégrale de la densité $\mathrm{Dirichlet}_W(\boldsymbol{\phi}_z\,; \mathbf{n}_z + \beta)$, égale à 1 par définition. On peut déduire par une méthode similaire que

$$
p(\mathbf{z}\,;\alpha) = \prod_{d=1}^{D} \frac{\mathrm{B}(\mathbf{n}_d + \alpha)}{\mathrm{B}(\alpha)} \tag{3.11}
$$

où $\mathbf{n}_d$ est le vecteur $(n_{dz})_{z=1}^{T}$ et n_{dz} désigne le nombre de mots attribués au thème z dans le document d.

Calcul de $p(z_{dn}|\mathbf{z}^{-(dn)}, \mathbf{w}\,;\alpha,\beta)$. Maintenant que nous avons calculé la probabilité jointe $p(\mathbf{w}, \mathbf{z}\,;\alpha,\beta)$, revenons à la probabilité centrale de l'échantillonneur de Gibbs marginalisé pour LDA : $p(z_{dn}|\mathbf{z}^{-(dn)}, \mathbf{w}\,;\alpha,\beta)$. Cette probabilité est obtenue de la manière suivante :

$$
\begin{aligned}
p(z_{dn} = z|\mathbf{z}^{-(dn)}, w_{dn} = w, \mathbf{w}^{-(dn)}\,;\alpha,\beta) &\propto \frac{p(\mathbf{w}|\mathbf{z}\,;\beta)}{p(\mathbf{w}^{-(dn)}|\mathbf{z}^{-(dn)}\,;\beta)} \cdot \frac{p(\mathbf{z}\,;\alpha)}{p(\mathbf{z}^{-(dn)}\,;\alpha)} \\
&= \frac{\mathrm{B}(\mathbf{n}_z + \beta)}{\mathrm{B}(\mathbf{n}_z^{-(dn)} + \beta)} \cdot \frac{\mathrm{B}(\mathbf{n}_d + \alpha)}{\mathrm{B}(\mathbf{n}_d^{-(dn)} + \alpha)} \\
&= \frac{\Gamma(n_{z\bullet}^{-(dn)} + W\beta)}{\Gamma(n_{z\bullet} + W\beta)} \cdot \frac{\Gamma(n_{zw} + \beta)}{\Gamma(n_{zw}^{-(dn)} + \beta)} \\
&\quad \cdot \frac{\Gamma(n_{d\bullet}^{-(dn)} + T\alpha)}{\Gamma(n_{d\bullet} + T\alpha)} \cdot \frac{\Gamma(n_{dz} + \alpha)}{\Gamma(n_{dz}^{-(dn)} + \alpha)} \\
&= \frac{n_{zw}^{-(dn)} + \beta}{n_{z\bullet}^{-(dn)} + W\beta} \cdot \frac{n_{dz}^{-(dn)} + \alpha}{n_{d\bullet}^{-(dn)} + T\alpha}
\end{aligned} \tag{3.12}
$$

où le symbole $\bullet$ en indice d'un compteur indique une sommation pour la variable à la place correspondante, c'est-à-dire $n_{z\bullet}^{-(dn)} = \sum_{w=1}^{W} n_{zw}^{-(dn)}$ et $n_{d\bullet}^{-(dn)} = \sum_{z=1}^{T} n_{dz}^{-(dn)}$. Le passage de la première à la deuxième ligne est effectué en remarquant que les ratios de fonctions Beta sont

6. Pour un vecteur $\mathbf{x}$ de taille N, $\mathrm{B}(\mathbf{x}) = \frac{\Gamma(x_1)\cdots\Gamma(x_N)}{\Gamma(x_1+\cdots+x_N)}$ avec $\Gamma(\cdot)$ la fonction Gamma d'Euler.

égaux à 1 sauf pour le thème et le document courants (z et d, respectivement). De manière similaire, les troisième et quatrième lignes sont obtenues en simplifiant les ratios de fonction Gamma pour des mots et thèmes autres que ceux courants (w et z, respectivement). Enfin, la dernière ligne est déduite en notant que les compteurs incluant le n^{e} mot du document d et les compteurs l'excluant ont une différence de 1 (par exemple, $n_{zw} = n_{zw}^{-(dn)} + 1$ car $z_{dn} = z$ et $w_{dn} = w$), puis en appliquant la propriété suivante de la fonction Gamma : $\Gamma(x+1) = x\,\Gamma(x)$ pour $x \in \mathbb{R}$.

Estimation de $\{\boldsymbol{\phi}_z\}_{z=1}^T$ et $\{\boldsymbol{\theta}_d\}_{d=1}^D$. À partir des attributions de thèmes pour chaque mot des documents de la collection, tirés selon la probabilité définie dans l'Équation (3.12), il est possible de calculer la probabilité postérieure des distributions de mots spécifiques aux thèmes $\{\boldsymbol{\phi}_z\}_{z=1}^T$:

$$p(\boldsymbol{\phi}_z|\mathbf{z},\mathbf{w}\,;\beta) \propto p(\mathbf{w}|\mathbf{z},\boldsymbol{\phi}_z)\,p(\boldsymbol{\phi}_z\,;\beta) \propto \prod_{w=1}^{W}(\phi_{zw})^{n_{zw}} \prod_{w=1}^{W}(\phi_{zw})^{\beta-1} \tag{3.13}$$

pour $z = 1, \ldots, T$. On reconnaît ici la densité $\text{Dirichlet}_W(\boldsymbol{\phi}_z\,;\mathbf{n}_z + \beta)$.[7] De la même manière, on trouve que la probabilité postérieure $p(\boldsymbol{\theta}_d|\mathbf{z}\,;\alpha)$ des distributions de thèmes spécifiques aux documents correspond à la densité $\text{Dirichlet}_T(\boldsymbol{\theta}_d\,;\mathbf{n}_d + \alpha)$, $d = 1, \ldots, D$. Ces lois postérieures peuvent alors être utilisées pour construire les estimateurs bayésiens suivants :

$$\hat{\phi}_{zw} = E\left[\phi_{zw}|\mathbf{z},\mathbf{w}\,;\beta\right] = \frac{n_{zw}+\beta}{n_{z\bullet}+W\beta} \qquad \text{avec} \qquad z = 1,\ldots,T,\ w = 1,\ldots,W\,; \tag{3.14}$$

$$\hat{\theta}_{dz} = E\left[\theta_{dz}|\mathbf{z}\,;\alpha\right] = \frac{n_{dz}+\alpha}{n_{d\bullet}+T\alpha} \qquad \text{avec} \qquad d = 1,\ldots,D,\ z = 1,\ldots,T\,; \tag{3.15}$$

où $E\left[\cdot|\cdot\right]$ désigne une espérance postérieure. Ces résultats découlent directement du fait que l'espérance de la i^{e} composante d'une variable aléatoire de dimension K distribuée suivant une loi $\text{Dirichlet}_K(\delta_1, \ldots, \delta_K)$ est égale à $\frac{\delta_i}{\sum_{i'=1}^K \delta_{i'}}$.

Résumé de l'algorithme complet. Ayant à disposition tous les éléments nécessaires à la construction d'un échantillonneur de Gibbs marginalisé pour LDA, nous résumons son déroulement dans l'Algorithme 3. L'algorithme prend en entrée le nombre d'itérations (ou de tirages) S, les valeurs des hyperparamètres[8] α et β des distributions $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$, respectivement, le nombre de thèmes T (fixés) et les mots apparaissant dans les documents de la collection $\mathbf{w} = \left\{\{w_{dn}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$. Dans un premier temps, les attributions de thèmes pour chaque mot de la collection sont initialisés aléatoirement en effectuant un tirage suivant une loi uniforme discrète sur les entiers entre 1 et T. Pour chaque attribution au thème z d'un mot de type w

7. En réalité, le fait que la probabilité postérieure d'une loi de Dirichlet soit également une loi de Dirichlet était prévisible : ce résultat découle de la relation particulière liant la loi de Dirichlet et la loi discrète (ainsi que la loi multinomiale), qui sont dites « conjuguées ».

8. Dans ce chapitre, nous considérons que les valeurs des hyperparamètres sont des constantes. Il est cependant possible de traiter ceux-ci comme des variables aléatoires et de les échantillonner, comme cela a été fait dans [Escobar et West, 1995; Newman *et al.*, 2009; Wallach *et al.*, 2009].

Algorithme 3 : Échantillonneur de Gibbs marginalisé pour LDA

Input : un nombre d'itérations S, les hyperparamètres α et β, le nombre de thèmes T, les mots des documents de la collection $\mathbf{w}$

Output : S tirages approximativement distribués selon $p(\mathbf{z}|\mathbf{w}\,;\alpha,\beta)$, S estimations des distributions $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$

// Initialisation aléatoire de $\mathbf{z}^{(0)}$

for $d \leftarrow 1$ **to** D, $n \leftarrow 1$ **to** N_d **do**

  $w \leftarrow w_{dn}$

  Tirer $z_{dn}^{(0)} \sim \text{Uniforme}(1, ..., T)$

  // Ajouter l'attribution du thème $z_{dn}^{(0)}$ aux compteurs

  $z \leftarrow z_{dn}^{(0)}$

  $n_{zw} \leftarrow n_{zw} + 1$

  $n_{dz} \leftarrow n_{dz} + 1$

end

// Construction de la chaîne de Markov pour obtenir les tirages $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$

for $s \leftarrow 1$ **to** S **do**

  // Tirage de $\mathbf{z}^{(s)}$

  for $d \leftarrow 1$ **to** D, $n \leftarrow 1$ **to** N_d **do**

    $w \leftarrow w_{dn}$

    // Retirer l'attribution de l'ancien thème $z_{dn}^{(s-1)}$ des compteurs

    $z \leftarrow z_{dn}^{(s-1)}$

    $n_{zw} \leftarrow n_{zw} - 1$ // $n_{zw}^{-(dn)}$

    $n_{dz} \leftarrow n_{dz} - 1$ // $n_{dz}^{-(dn)}$

    Tirer $z_{dn}^{(s)} \sim \dfrac{1}{\mathcal{Z}} \displaystyle\sum_{z=1}^{T} \frac{n_{zw}+\beta}{n_{z\bullet}+W\beta} \frac{n_{dz}+\alpha}{n_{d\bullet}+T\alpha}\, \delta_z(\cdot)$

    // Ajouter l'attribution du nouveau thème $z_{dn}^{(s)}$ aux compteurs

    $z' \leftarrow z_{dn}^{(s)}$

    $n_{z'w} \leftarrow n_{z'w} + 1$

    $n_{dz'} \leftarrow n_{dz'} + 1$

  end

  // Estimation de $\boldsymbol{\theta}$ et $\boldsymbol{\phi}$

  for $z \leftarrow 1$ **to** T, $w \leftarrow 1$ **to** W **do**

    $\hat{\phi}_{zw}^{(s)} \leftarrow \dfrac{n_{zw}+\beta}{n_{z\bullet}+W\beta}$

  end

  for $d \leftarrow 1$ **to** D, $z \leftarrow 1$ **to** T **do**

    $\hat{\theta}_{dz}^{(s)} \leftarrow \dfrac{n_{dz}+\alpha}{n_{d\bullet}+T\alpha}$

  end

end

return $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\theta}}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\phi}}^{(s)}\right\}_{s=1}^{S}$

dans le document d, les compteurs n_{zw} et n_{dz} sont incrémentés. Ensuite, les tirages $\mathbf{z}^{(s)}$ pour chaque échantillon $s = 1, ..., S$ sont générés successivement. Pour le n^{e} mot du document d, de type w, le thème z précédemment attribué (c'est-à-dire pour le tirage $s-1$) est « retiré » : les compteurs n_{zw} et n_{dz} sont décrémentés pour exclure le thème de ce mot. Après décrémentation, ces compteurs correspondent en réalité respectivement à $n_{zw}^{-(dn)}$ et $n_{dz}^{-(dn)}$. Le nouveau thème z' attribué à $z_{dn}^{(s)}$ est alors tiré suivant la loi discrète dont la densité a été établie dans l'Équation (3.12) : $\frac{1}{\mathcal{Z}} \sum_{z=1}^{T} \frac{n_{zw}+\beta}{n_{z\cdot}+W\beta} \frac{n_{dz}+\alpha}{n_{d\cdot}+T\alpha} \delta_z(\cdot)$, où $\delta_z(\cdot)$ désigne la masse de Dirac centrée en z et $\mathcal{Z}$ désigne la constante de normalisation. Après tirage du thème z' nouvellement attribué, les compteurs $n_{z'w}$ et $n_{dz'}$ sont incrémentés. Enfin, après avoir attribué un thème à chaque mot de chaque document dans la collection pour le tirage courant s, les distributions de mots spécifiques aux thèmes $\boldsymbol{\phi}$ et les distributions de thèmes spécifiques aux documents $\boldsymbol{\theta}$ sont estimées en appliquant les Équations (3.14) et (3.15), respectivement. Après S itérations, les tirages $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$ ainsi que les estimateurs $\left\{\hat{\boldsymbol{\theta}}^{(s)}\right\}_{s=1}^{S}$ et $\left\{\hat{\boldsymbol{\phi}}^{(s)}\right\}_{s=1}^{S}$ sont renvoyés.

3.3.2 Inférence variationnelle

Dans la littérature sur les modèles thématiques, des méthodes d'inférence postérieure approchées autres que l'échantillonnage de Gibbs ont également été explorées. La plus commune d'entre elles est l'inférence variationnelle [Jordan *et al.*, 1999]. Étant donné que nous n'utilisons pas cette méthode dans nos contributions, nous présentons le principe général de l'inférence variationnelle et esquissons simplement l'application de l'inférence variationnelle à LDA, en laissant le lecteur se référer à [Blei *et al.*, 2001, 2003] pour une description plus détaillée de la procédure complète.

3.3.2.1 Principe général

Soient un ensemble de variables aléatoires observées $\mathbf{x} = \{x_i\}_{i=1}^{N}$ et un ensemble de variables aléatoires latentes $\mathbf{z} = \{z_j\}_{j=1}^{P}$. L'objectif de l'inférence variationnelle est de fournir une approximation de la loi postérieure $p(\mathbf{z}|\mathbf{x})$, en considérant que cette loi ne peut être calculée de manière exacte. Le principe de la méthode variationnelle est alors d'approcher $p(\mathbf{z}|\mathbf{x})$ par une loi $q_{\boldsymbol{\omega}}(\mathbf{z})$ plus simple dépendant d'un ensemble de paramètres libres $\boldsymbol{\omega}$ que l'on va devoir déterminer. Trouver une loi $q_{\boldsymbol{\omega}}(\mathbf{z})$ qui est « proche » de $p(\mathbf{z}|\mathbf{x})$ revient à identifier $\boldsymbol{\omega}$ tel que la distance entre ces deux lois soit minimisée. Une distance communément utilisée pour les lois de probabilité est la divergence de Kullback-Leibler :

$$\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\omega}} \parallel p) = \int_{\mathbf{z}} q_{\boldsymbol{\omega}}(\mathbf{z}) \log \frac{q_{\boldsymbol{\omega}}(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} \,\mathrm{d}\mathbf{z}. \tag{3.16}$$

En utilisant le fait que $p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{z},\mathbf{x})}{p(\mathbf{x})}$ et $\int_{\mathbf{z}} q_{\boldsymbol{\omega}}(\mathbf{z})\mathrm{d}\mathbf{z} = 1$, on obtient :

$$\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\omega}} \parallel p) = \log p(\mathbf{x}) + \int_{\mathbf{z}} q_{\boldsymbol{\omega}}(\mathbf{z}) \log \frac{q_{\boldsymbol{\omega}}(\mathbf{z})}{p(\mathbf{z},\mathbf{x})} \,\mathrm{d}\mathbf{z} \tag{3.17}$$

$$
\begin{aligned}
&= \log p(\mathbf{x}) + E_{q_{\boldsymbol{\omega}}}\left[\log q_{\boldsymbol{\omega}}(\mathbf{z})\right] - E_{q_{\boldsymbol{\omega}}}\left[\log p(\mathbf{z},\mathbf{x})\right] \\
&= \log p(\mathbf{x}) - \mathcal{L}(q_{\boldsymbol{\omega}}).
\end{aligned}
$$

où $\mathcal{L}(q_{\boldsymbol{\omega}}) = E_{q_{\boldsymbol{\omega}}}\left[\log p(\mathbf{z},\mathbf{x})\right] - E_{q_{\boldsymbol{\omega}}}\left[\log q_{\boldsymbol{\omega}}(\mathbf{z})\right]$ est une quantité connue sous le nom d'« énergie variationnelle libre négative » (*negative variational free energy*) ou d'ELBO (*evidence lower bound*) dans la littérature.

Minimiser la divergence de Kullback-Leibler entre $q_{\boldsymbol{\omega}}$ et p est donc équivalent au problème d'optimisation suivant :

$$
\max_{\boldsymbol{\omega}} \mathcal{L}(q_{\boldsymbol{\omega}}) = E_{q_{\boldsymbol{\omega}}}\left[\log p(\mathbf{z},\mathbf{x})\right] - E_{q_{\boldsymbol{\omega}}}\left[\log q_{\boldsymbol{\omega}}(\mathbf{z})\right]. \tag{3.18}
$$

L'avantage d'optimiser $\mathcal{L}(q_{\boldsymbol{\omega}})$ plutôt que $\mathrm{D}_{\mathrm{KL}}(q_{\boldsymbol{\omega}} \parallel p)$ provient du fait que $\mathcal{L}(q_{\boldsymbol{\omega}})$ dépend de la probabilité jointe $p(\mathbf{z},\mathbf{x})$, qui se déduit immédiatement du modèle graphique, et non de la probabilité postérieure $p(\mathbf{z}|\mathbf{x})$, que l'on cherche justement à approcher.

Afin de rendre le calcul de $\mathcal{L}(q_{\boldsymbol{\omega}})$ analytiquement possible, la loi $q_{\boldsymbol{\omega}}$ est généralement choisie dans une famille de distributions simples. Une telle famille de distributions est obtenue en appliquant l'hypothèse dite de « champ moyen » (*mean field assumption*) selon laquelle $q_{\boldsymbol{\omega}}(\mathbf{z})$ peut être factorisée suivant une partition $\{\mathbf{z}_k\}_{k=1}^{Q}$ des variables latentes $\{z_j\}_{j=1}^{P}$:

$$
q_{\boldsymbol{\omega}}(\mathbf{z}) = \prod_{k=1}^{Q} q_{\omega_k}(\mathbf{z}_k). \tag{3.19}
$$

avec $\boldsymbol{\omega} = \{\omega_k\}_{k=1}^{Q}$. Suivant cette hypothèse, il est alors possible de montrer [Bishop, 2006] qu'un maximum local $q^*_{\boldsymbol{\omega}}(\mathbf{z})$ de $\mathcal{L}(q_{\boldsymbol{\omega}})$ est atteint en calculant itérativement jusqu'à convergence chaque facteur $q^*_{\omega_k}(\mathbf{z}_k)$, $k = 1, \ldots, Q$, tel que :

$$
q^*_{\omega_k}(\mathbf{z}_k) \propto \exp\left\{ \prod_{\substack{\ell=1 \\ \ell \neq k}}^{Q} \int_{\mathbf{z}_\ell} q^*_{\omega_\ell}(\mathbf{z}_\ell) \log p(\mathbf{z},\mathbf{x}) \, \mathrm{d}\mathbf{z}_\ell \right\}. \tag{3.20}
$$

Cela permet ainsi d'obtenir une approximation $q^*_{\boldsymbol{\omega}}(\mathbf{z})$ du postérieur exact $p(\mathbf{z}|\mathbf{x})$.

3.3.2.2 Application à LDA

Dans le cas du modèle LDA, rappelons que les variables observées sont les mots $\mathbf{w} = \left\{\{w_{dn}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$ et les variables latentes sont les attributions de thèmes $\mathbf{z} = \left\{\{z_{dn}\}_{n=1}^{N_d}\right\}_{d=1}^{D}$, les distributions de thèmes spécifiques aux documents $\boldsymbol{\theta} = \{\boldsymbol{\theta}_d\}_{d=1}^{D}$ et les distributions de mots spécifiques aux thèmes $\boldsymbol{\phi} = \{\boldsymbol{\phi}_z\}_{z=1}^{T}$. La loi $q_{\boldsymbol{\omega}}(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi})$ approchant le postérieur $p(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\phi}|\mathbf{w}\,;\alpha,\beta)$ peut être définie selon la factorisation complète suivante [Blei *et al.*, 2001,

2003] :

$$q_{\boldsymbol{\omega}}(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \left\{ \prod_{d=1}^{D} \prod_{n=1}^{N_d} q_{\boldsymbol{\pi}_d}(z_{dn}) \right\} \left\{ \prod_{d=1}^{D} q_{\gamma_d}(\boldsymbol{\theta}_d) \right\} \left\{ \prod_{z=1}^{T} q_{\eta_z}(\boldsymbol{\phi}_z) \right\} \tag{3.21}$$

$$= \left\{ \prod_{d=1}^{D} \prod_{n=1}^{N_d} \pi_{dz_{dn}} \right\} \left\{ \prod_{d=1}^{D} \text{Dirichlet}_T(\boldsymbol{\theta}_d \,;\, \gamma_d) \right\} \left\{ \prod_{z=1}^{T} \text{Dirichlet}_W(\boldsymbol{\phi}_z \,;\, \eta_z) \right\}$$

où $\boldsymbol{\omega} = \left\{ \{\boldsymbol{\pi}_d\}_{d=1}^{D}, \{\gamma_d\}_{d=1}^{D}, \{\eta_z\}_{z=1}^{T} \right\}$ correspond à l'ensemble des paramètres libres à déterminer. Blei *et al.* [2001, 2003] ont montré que les paramètres $\boldsymbol{\omega}$ ainsi que les hyperparamètres α et β peuvent être estimés à travers un algorithme de type espérance-maximisation (*expectation-maximization*). Nous laissons le lecteur se référer à ce papier pour plus de détails sur cette procédure, qui dépasse le cadre de cette thèse.

Les différentes méthodes d'inférence postérieure que nous avons décrites dans la Section 3.3 sont facilement généralisables aux modèles thématiques étendant LDA – ce qui explique la popularité de ce type de modèles. Après avoir défini une extension de LDA et réalisé son inférence, il est pertinent de comparer ce nouveau modèle à LDA ou à d'autres variantes afin d'identifier dans quelle mesure les différentes approches fournissent une modélisation appropriée des données observées. Dans la Section 3.4, nous détaillons les différentes méthodes d'évaluation quantitative qui ont été proposées dans la littérature.

3.4 Évaluation

L'évaluation des modèles thématiques a été l'objet de multiples travaux dans la littérature [Chang *et al.*, 2009; Lau *et al.*, 2014; Mimno *et al.*, 2011; Newman *et al.*, 2010; Röder *et al.*, 2015] et demeure un sujet de recherche actif. Traditionnellement, les modèles thématiques étaient évalués en terme de perplexité ou de log-vraisemblance, mesures statistiques de la capacité d'un modèle à généraliser à partir des données d'apprentissage. Il a cependant été démontré dans [Chang *et al.*, 2009] que ces mesures ne sont pas positivement corrélées avec la cohérence des thèmes perçue par les humains. Par conséquent, des métriques capables de calculer automatiquement la cohérence thématique d'un modèle ont par la suite été introduites dans [Lau *et al.*, 2014; Mimno *et al.*, 2011; Newman *et al.*, 2010; Röder *et al.*, 2015].

Dans le reste de cette section, nous présentons les principales méthodes d'évaluation appliquées aux modèles thématiques dans la littérature. Nous détaillons la perplexité et les différentes mesures de cohérence thématique dans les Sections 3.4.1 et 3.4.2, respectivement. Puis, dans la Section 3.4.3, nous abordons l'évaluation de modèles thématiques basée sur des tâches externes (par exemple, regroupement ou classification) plutôt que sur des métriques inhérentes aux modèles thématiques.

3.4.1 Perplexité

La perplexité est une mesure statistique utilisée pour évaluer la capacité de généralisation d'un modèle de langue. D'un point de vue théorique, cette mesure peut être interprétée comme l'inverse de la moyenne géométrique de la vraisemblance sur l'ensemble des mots de la collection. Après entraînement d'un modèle sur un ensemble d'apprentissage, la perplexité de ce modèle est ensuite mesurée sur un ensemble de test distinct. Une perplexité basse sur l'ensemble de test (équivalant à une vraisemblance élevée) signifie que le modèle est peu « surpris » par ces nouvelles données, ce qui indique une meilleure capacité du modèle à généraliser à partir des données d'apprentissage.

Dans le cas de LDA (et de ses variantes), la perplexité n'est pas calculable de manière exacte mais elle peut être approchée comme suit pour un ensemble de test $\mathbf{w}^{\text{test}}$:

$$\begin{aligned} \text{perp}(\mathbf{w}^{\text{test}}) &= \exp\left\{-\frac{\sum_{d=1}^{D^{\text{test}}}\sum_{n=1}^{N_d}\log\hat{p}(w_{dn})}{\sum_{d=1}^{\text{test}}N_d}\right\} \\ &= \exp\left\{-\frac{\sum_{d=1}^{D^{\text{test}}}\sum_{n=1}^{N_d}\log\left(\sum_{z=1}^{T}\hat{\theta}_{dz}\hat{\phi}_{zw_{dn}}\right)}{\sum_{d=1}^{D}N_d}\right\} \end{aligned} \tag{3.22}$$

$\hat{\boldsymbol{\theta}}$ et $\hat{\boldsymbol{\phi}}$ correspondent respectivement aux estimateurs des distributions de thèmes spécifiques aux documents et aux estimateurs des distributions de mots spécifiques aux thèmes. Ces quantités sont similaires à celles obtenues en sortie de l'échantillonneur de Gibbs marginalisé décrit dans l'Algorithme 3.

On note toutefois que l'estimateur $\hat{\theta}_{dz}$ porte sur un document de l'ensemble de test. Or le modèle (et ses estimateurs) devrait être entraîné sur l'ensemble d'apprentissage et non sur l'ensemble de test. Pour pallier ce problème, différentes stratégies ont été employées dans la littérature. Heinrich [2008] a proposé une procédure en deux étapes. LDA est tout d'abord entraîné sur l'ensemble d'apprentissage par un échantillonneur de Gibbs marginalisé pendant S itérations. À partir du dernier état S obtenu, l'échantillonneur va ensuite parcourir uniquement l'ensemble de test pour un nombre S' d'itérations, en conservant les compteurs associés à l'ensemble d'apprentissage mais sans mettre ceux-ci à jour. À la fin de cette procédure, les estimateurs $\hat{\phi}_{zw}$ ($z = 1, ..., T$, $w = 1, ..., W$) et $\hat{\theta}_{dz}$ ($d = 1, ..., D^{\text{test}}$, $z = 1, ..., T$) utilisés pour calculer la perplexité sur l'ensemble de test correspondent aux quantités suivantes :

$$\hat{\phi}_{zw} = E\left[\phi_{zw}\middle|\mathbf{z}^{\text{train}}, \mathbf{z}^{\text{test}}, \mathbf{w}^{\text{train}}, \mathbf{w}^{\text{test}}\right] = \frac{n_{zw}^{\text{train}} + n_{zw}^{\text{test}} + \beta}{n_{z\bullet}^{\text{train}} + n_{z\bullet}^{\text{test}} + W\beta}\,; \tag{3.23}$$

$$\hat{\theta}_{dz} = E\left[\theta_{dz}\middle|\mathbf{z}^{\text{test}}\right] = \frac{n_{dz}^{\text{test}} + \alpha}{n_{d\bullet}^{\text{test}} + T\alpha}\,; \tag{3.24}$$

où les mentions « train » et « test » en exposant indiquent que les variables et compteurs concernés dépendent respectivement de l'ensemble d'apprentissage et de l'ensemble de test.

Une autre méthode employée dans [Asuncion *et al.*, 2009; Newman *et al.*, 2009] consiste à découper en deux chaque document de test : une moitié est ajoutée à un ensemble dit de

fold-in, l'autre moitié est conservée dans l'ensemble de test – et sera utilisée pour le calcul de la perplexité. La procédure adoptée est ensuite conduite en deux étapes. Dans un premier temps, l'échantillonneur de Gibbs marginalisé est appliqué à l'ensemble d'apprentissage, permettant d'obtenir les estimateurs $\hat{\boldsymbol{\phi}}$. Dans un second temps, un autre échantillonneur de Gibbs est appris uniquement sur les documents *fold-in* en utilisant les estimateurs $\hat{\boldsymbol{\phi}}$ appris précédemment et sans mettre ceux-ci à jour. De cette manière, le second échantillonneur de Gibbs appliqué à l'ensemble *fold-in* (contenant les moitiés de documents de test) fournit naturellement un estimateur $\hat{\boldsymbol{\theta}}_d$ pour chaque document de test $d = 1, ..., D^{\text{test}}$. La perplexité est alors testée sur les moitiés restantes de documents de test à partir des estimateurs $\hat{\boldsymbol{\theta}}$ et $\hat{\boldsymbol{\phi}}$ obtenus à l'issue des deux phases de la procédure. L'avantage de cette méthode repose sur le fait que les mots sur lesquels la perplexité est testée ne sont jamais utilisés dans la phase d'apprentissage [9] – contrairement à la méthode d'Heinrich.

Malgré sa popularité, la perplexité a été l'objet de plusieurs critiques en raison de sa nature purement statistique. En particulier, Chang *et al.* [2009] ont observé qu'un modèle avec une perplexité plus basse qu'un autre avait en réalité tendance à renvoyer des thèmes moins cohérents que ce dernier. Pour cette raison, plusieurs chercheurs ont par la suite développé des mesures de cohérence thèmatique, que nous détaillons dans la Section 3.4.2.

3.4.2 Cohérence thématique

Après l'inférence du modèle LDA, chaque thème $z = 1, ..., T$ est associé à une distribution de mots $\{\hat{\phi}_{zw}\}_{w=1}^{W}$ spécifique à ce thème. Une manière de représenter le thème z est de sélectionner les N mots les plus probables de $\{\hat{\phi}_{zw}\}_{w=1}^{W}$, que nous dénotons par $\{w_{zi}\}_{i=1}^{N}$. Formellement, cela signifie que $\hat{\phi}_{zw_{z1}} \geq ... \geq \hat{\phi}_{zw_{zN}} \geq \hat{\phi}_{zw'}$ pour tout $w' \notin \{w_{zi}\}_{i=1}^{N}$. Ainsi un thème peut être considéré comme étant de « bonne qualité » si ses N mots les plus probables forment un ensemble cohérent. Les différentes mesures de cohérence thématique proposées dans la littérature [Lau *et al.*, 2014; Mimno *et al.*, 2011; Newman *et al.*, 2009; Röder *et al.*, 2015] sont basées sur le phénomène de co-occurrence : deux mots ont plus de chance de refléter le même thème s'ils apparaissent fréquemment ensemble (par exemple, dans le même document ou la même phrase). Le corpus d'où sont extraites les informations de co-occurrence est généralement différent de la collection sur laquelle le modèle thématique est appris. En effet, afin de disposer d'un nombre important de co-occurrences, ce corpus est choisi de grande taille, typiquement de l'ordre du million de documents (par exemple, un sous-ensemble des articles Wikipédia ou des articles d'actualité publiés par le New York Times pendant une décennie [Lau *et al.*, 2014]).

Détaillons maintenant les principales mesures de cohérence thématique utilisées dans la littérature. La mesure développée par Mimno *et al.* [2011] consiste à calculer la log-probabilité

9. Cela est souhaitable pour déterminer avec exactitude dans quelle mesure le modèle est « surpris » par l'ensemble de test.

conditionnelle des mots $\{w_{zi}\}_{i=1}^{N}$ de la manière suivante :

$$\mathrm{C}_{\mathrm{Mimno}}(z) = \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{p(w_{zi}, w_{zj}) + \epsilon}{p(w_{zj})}. \tag{3.25}$$

La probabilité jointe $p(w_{zi}, w_{zj})$ est estimée par $\frac{D(w_{zi}, w_{zj})}{D}$ où $D(w_{zi}, w_{zj})$ dénote le nombre de documents dans lesquels w_{zi} et w_{zj} apparaissent [10]. La probabilité marginale $p(w_{zj})$ est définie comme $\frac{D(w_{zj})}{D}$ où $D(w_{zj})$ désigne le nombre de documents dans lesquels le mot w_{zj} apparaît. La constante réelle positive ϵ (typiquement choisie telle que $\epsilon \leq 1$) est introduite pour éviter l'apparition de $\log(0)$.

La variante symétrisée de cette mesure, basée sur l'information mutuelle ponctuelle (PMI), a été proposée par Newman *et al.* [2009] :

$$\mathrm{C}_{\mathrm{Newman}}(z) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \log \frac{p(w_{zi}, w_{zj}) + \epsilon}{p(w_{zi})\, p(w_{zj})}. \tag{3.26}$$

Lau *et al.* [2014] ont par la suite étudié la version normalisée de $\mathrm{C}_{\mathrm{Newman}}$, définie à partir de l'information mutuelle ponctuelle normalisée (NPMI) :

$$\mathrm{C}_{\mathrm{Lau}}(z) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{-\log \frac{p(w_{zi}, w_{zj}) + \epsilon}{p(w_{zi})\, p(w_{zj})}}{\log \{p(w_{zi}, w_{zj}) + \epsilon\}}. \tag{3.27}$$

Afin de comparer ces mesures et d'en proposer de nouvelles, Röder *et al.* [2015] ont exploré de manière systématique l'espace des mesures de cohérence thématique.

Les différentes mesures d'évaluation que nous avons mentionnées jusqu'ici – la perplexité et les mesures de cohérence thématique – se basent directement sur le modèle thématique appris lors de la phase d'inférence. Plutôt qu'évaluer le modèle de manière intrinsèque, il est possible d'appliquer ce modèle à une tâche externe (par exemple, regroupement ou classification) et d'en évaluer les performances dans le cadre de cette tâche. Nous détaillons cette approche d'évaluation alternative dans la Section 3.4.3.

3.4.3 Évaluation basée sur des tâches externes

En tant que méthodes exploratoires, les modèles thématiques peuvent être appliqués à de multiples tâches : recherche d'information [Wei et Croft, 2006], classification [Li *et al.*, 2016], regroupement [Shafiei et Muios, 2006], systèmes de recommandation [Barbieri *et al.*, 2014], génération de résumés [Tang *et al.*, 2009], etc. Ainsi, au lieu de comparer la perplexité ou la

10. Le calcul du nombre de co-occurrences peut alternativement être basé sur la présence des mots w_{zi} et w_{zj} dans une fenêtre de taille définie que l'on fait « glisser » pour couvrir l'intégralité de chaque document. Les informations de co-occurrence ainsi obtenues sont plus précises, en particulier dans le cas des documents longs.

cohérence thématique de différents modèles thématiques, il est possible d'évaluer ceux-ci par l'intermédiaire des tâches susmentionnées.

Supposons par exemple que l'on souhaite comparer deux modèles thématiques $\mathcal{M}$ et $\mathcal{M}'$ dans le cadre de la classification de documents. Après inférence de ces modèles, on obtient pour chaque document d deux représentations $\hat{\boldsymbol{\theta}}_d$ et $\hat{\boldsymbol{\theta}}'_d$, respectivement, sous la forme de vecteurs de dimension T. La performance d'un classifieur quelconque (par exemple, SVM) entraîné indépendamment avec chacune de ces deux représentations comme traits indiquera alors quelle représentation des données est la plus discriminante et, par conséquent, quel modèle thématique est le plus performant.

3.5 Conclusion

Dans ce chapitre, nous avons présenté les notions de base nécessaires à la compréhension et à l'application du modèle thématique LDA et de ses extensions. Nous avons tout d'abord défini LDA à partir de son histoire générative et de sa représentation sous forme de modèle graphique. La probabilité postérieure de LDA ne pouvant être calculée de manière exacte, nous avons ensuite détaillé les principales méthodes d'inférence postérieure approchée appliquées aux modèles thématiques : échantillonneur de Gibbs et inférence variationnelle. Enfin, nous avons abordé la question de l'évaluation des modèles thématiques en décrivant en particulier les mesures de perplexité et de cohérence thématique communément utilisées dans la littérature.

Fort de cette introduction au formalisme propre aux modèles thématiques, nous décrivons dans le reste de cette thèse (Chapitres 4 et 5) nos contributions, qui consistent à étendre le modèle LDA pour intégrer la notion de point de vue.

Deuxième partie

Contributions à la découverte de points de vue sur le Web

Chapitre 4

Découverte de points de vue dans les documents textuels

Sommaire

Dans ce chapitre, nous décrivons notre première contribution à la fouille de points de vue [Thonet *et al.*, 2016], dans laquelle nous nous focalisons sur la modélisation des points de vue dans les documents textuels sans métadonnées disponibles (telles que des informations portant sur l'auteur, son activité en ligne ou son réseau d'amis). Nous définissons en particulier la tâche de *découverte des points de vue et des opinions* (*viewpoint and opinion discovery*), qui consiste à analyser une collection de documents afin d'identifier le point de vue de chaque document, les thèmes mentionnés par chaque document et les opinions [1] spécifiques aux points de vue pour chaque thème. Pour traiter ce problème, nous proposons dans la Section 4.1 le modèle VODUM (*viewpoint and opinion discovery unification model*), une approche non supervisée permettant la modélisation conjointe des points de vue et des thèmes en différenciant mots d'opinion (spécifiques aux points de vue et aux thèmes) et mots thématiques (spécifiques aux thèmes uniquement et neutres vis-à-vis des différents points de vue). À travers VODUM, nous étudions dans quelle mesure les *parties de discours* peuvent être exploitées pour distinguer les mots d'opinion et les mots thématiques dans un cadre non supervisé. Dans la Section 4.2, VODUM est ensuite évalué quantitativement et qualitativement sur la collection Bitterlemons, comparé aux modèles de l'état de l'art ainsi qu'à des versions dégénérées de notre modèle pour étudier l'apport de chaque spécificité de VODUM de manière isolée. Enfin, la Section 4.3 évoque les extensions possibles de VODUM et conclut le chapitre.

1. Une opinion désigne ici le choix de mots associé à un point de vue et à un thème.

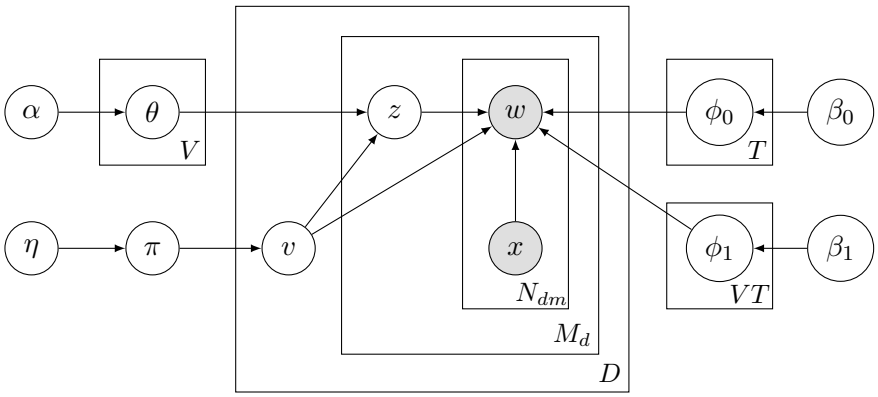


FIGURE 4.1 – Représentation de VODUM sous forme de modèle graphique.

4.1 VODUM : un modèle unifiant la découverte des thèmes, des opinions et des points de vue

VODUM (*viewpoint and opinion discovery unification model*) est un modèle probabiliste basé sur LDA [Blei *et al.*, 2001, 2003]. VODUM modélise de manière conjointe les points de vue et les thèmes, afin d'identifier les mots thématiques (associés à un thème et indépendant des points de vue) et les mots d'opinion (spécifiques à la fois à un thème et à un point de vue). Dans la Section 4.1.1, nous apportons une définition détaillée de notre modèle, puis nous décrivons son inférence dans la Section 4.1.2.

4.1.1 Description du modèle

Le modèle graphique représentant VODUM et les notations utilisés dans ce chapitre sont fournis dans la Figure 4.1 et le Tableau 4.1, respectivement. Avant de donner l'histoire générative de VODUM, nous décrivons les spécificités de notre modèle.

Distinction entre mots thématiques et mots d'opinion. Dans notre modèle VODUM, les mots thématiques et les mots d'opinion sont partitionnés en fonction de leur partie de discours, en accord avec différents travaux portant sur la fouille d'opinion [Liu *et al.*, 2005; Turney, 2002] et la fouille de points de vue [Fang *et al.*, 2012; Joshi *et al.*, 2016], que nous avons détaillés dans les Chapitres 1 et 2, respectivement. Nous avons ainsi supposé que les noms propres et communs sont des mots thématiques, et les adjectifs, verbes et adverbes des mots d'opinion [2]. L'étape de pré-traitement permettant d'obtenir les étiquettes de parties du discours sera décrite plus avant dans la Section 4.2.1.2. Afin de séparer les mots thématiques et les mots d'opinion, nous utilisons ce que nous nommons des *catégories de partie du discours*, associées à chaque mot d'une collection de documents. Elles sont représentées par des variables

2. Bien que cette séparation puisse sembler grossière, nous soulignons qu'une définition plus précise des mots thématiques et d'opinion pourrait être adoptée (par exemple, en exploitant des lexiques de subjectivité) sans que cela ne requiert de modification sur notre modèle.

Tableau 4.1 – Notations adoptées pour décrire VODUM.

Symbole	Définition
D, M_d, N_{dm}	Nombre de documents dans la collection, nombre de phrases dans le document d, nombre de mots dans la phrase m du document d, respectivement.
W	Taille du vocabulaire.
W_0, W_1	Nombre de mots thématiques et de mots d'opinion dans le vocabulaire, respectivement.
T, V	Nombre de thèmes et de points de vue, respectivement.
$\mathcal{W}_0$, $\mathcal{W}_1$	Ensemble des mots thématiques et des mots d'opinion dans le vocabulaire, respectivement.
w_{dmn}	Le n^{e} mot de la phrase m du document d.
x_{dmn}	La catégorie de partie du discours (0 ou 1) du n^{e} mot de la phrase m du document d.
z_{dm}	Le thème attribué à la phrase m du document d.
v_d	Le point de vue attribué au document d.
$\mathbf{w}$, $\mathbf{x}$, $\mathbf{z}$, $\mathbf{v}$	Vecteur de tous les mots, catégories de partie du discours, attributions de thèmes et attributions de points de vue, respectivement.
$\boldsymbol{\phi_0}$	Matrice de taille $T \times W$ contenant les matrices de distributions de mots thématiques (dépendantes du thème et indépendantes du point de vue).
$\boldsymbol{\phi_1}$	Matrice de taille $V \times T \times W$ contenant les distributions de mots d'opinions (dépendantes du point de vue et du thème).
$\boldsymbol{\theta}$	Matrice de taille $V \times T$ contenant les distributions de thèmes spécifiques aux points de vue.
$\boldsymbol{\pi}$	Matrice de taille $1 \times V$ contenant la distribution de points de vue.
β_0, β_1, α, η	Paramètre du *prior* de Dirichlet symétrique sur ϕ_0, ϕ_1, θ et π, respectivement.
n_v	Nombre de documents attribués au point de vue v.
n_{vz}	Nombre de phrases attribuées au point de vue v et au thème z.
n_{0zw}	Nombre d'occurrences du mot thématique w attribuées au thème z.
n_{1vzw}	Nombre d'occurrences du mot d'opinion w attribuées au point de vue v et au thème z.

observées $\mathbf{x}$ qui prennent la valeur 0 pour les mots thématiques et la valeur 1 pour les mots d'opinion. Les mots thématiques et les mots d'opinion sont considérés comme étant tirés suivant des distributions $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$, respectivement.

Attributions de thèmes au niveau des phrases. La plupart des modèles thématiques définissent les attributions de thèmes au niveau des mots, dans la lignée de LDA. Nous faisons cependant l'hypothèse que l'utilisation d'attributions de thèmes positionnés au niveau des phrases (dénotés par $\mathbf{z}$ dans VODUM) permet de mieux capturer la dépendance entre les mots thématiques et d'opinions employés dans une phrase et le thème de cette phrase. Par conséquent, les thèmes induits par les distributions de mots thématiques $\boldsymbol{\phi}_0$ et les distributions de mots d'opinions $\boldsymbol{\phi}_1$ ont plus de chance d'être alignés.

Attributions de points de vue au niveau des documents. Dans VODUM, les variables $\mathbf{v}$ dénotant les attributions de points de vue sont définies au niveau des documents et tirés suivant une distribution $\boldsymbol{\pi}$. Dans les approches similaires à la nôtre, les points de vue sont attribués aux mots [Paul et Girju, 2010; Paul *et al.*, 2010] ou aux auteurs [Qiu et Jiang, 2013; Qiu *et al.*, 2013b]. Bien qu'il soit raisonnable de supposer qu'un auteur écrit tous ses documents suivant le même point de vue, l'identité de l'auteur d'un document n'est pas toujours connue. D'autre part, attribuer à chaque mot d'un document des points de vue potentiellement différents paraît peu judicieux. En effet, si un document contient fréquemment un mélange de thèmes, il semble toutefois moins probable que l'auteur du document change de point de vue en cours de rédaction. Nous avons par conséquent opté pour des points de vue positionnés au niveau des documents.

Distributions de thèmes spécifiques aux points de vue. Nous avons adopté des distributions de thèmes, dénotées par $\boldsymbol{\theta}$, qui sont spécifiques aux points de vue plutôt que spécifiques aux documents – cette dernière modélisation étant plus courante dans la littérature des modèles thématiques [Blei *et al.*, 2001, 2003; Paul et Girju, 2010; Trabelsi et Zaïane, 2014]. Notre choix est motivé par l'observation faite dans [Qiu et Jiang, 2013] selon laquelle différents points de vue mettent en exergue différents thèmes. Par exemple, les opposants au « mariage pour tous » mentionneront plus vraisemblablement le thème de la religion que le ferait le bord partisan.

Similairement à LDA et à la plupart de ses extensions, les attributions de thèmes $\mathbf{z}$, de points de vue $\mathbf{v}$ et les mots $\mathbf{w}$ sont tirés suivant des lois discrètes paramétrisées par $\boldsymbol{\theta}$, $\boldsymbol{\pi}$ et $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$, respectivement. Un *prior* de Dirichlet symétrique est placé sur ces distributions avec pour hyperparamètre α, η, β_0 et β_1, respectivement.

Après avoir introduit les différentes spécificités de VODUM, nous pouvons maintenant décrire l'histoire générative de notre modèle comme suit. Chaque document de la collection à générer se voit tout d'abord attribuer un point de vue, reflétant le point de vue de l'auteur. L'auteur attribue ensuite à chaque phrase le thème qu'il va aborder – choix effectué en fonction de son point de vue. Puis, pour chaque phrase, il choisit un ensemble de mots thématiques décrivant le thème sélectionné, ainsi qu'un ensemble de mots d'opinion qui expriment son point de vue vis-à-vis de ce thème. Formellement, la génération d'une collection de documents telle qu'expliquée par VODUM se déroule de la manière suivante :

1. Tirer une distribution de points de vue $\boldsymbol{\pi} \sim \text{Dirichlet}_V(\eta)$;
2. Tirer une distribution de mots thématiques indépendants du point de vue $\boldsymbol{\phi}_{0z} \sim \text{Dirichlet}_W(\beta_0)$ pour chaque thème $z = 1, ..., T$;
3. Tirer une distribution de mots d'opinions dépendants du point de vue $\boldsymbol{\phi}_{1vz} \sim \text{Dirichlet}_W(\beta_1)$ pour chaque point de vue $v = 1, ..., V$ et chaque thème $z = 1, ..., T$;
4. Tirer une distribution de thèmes $\boldsymbol{\theta}_v \sim \text{Dirichlet}_T(\alpha)$ pour chaque point de vue $v = 1, ..., V$;
5. Pour chaque document $d = 1, ..., D$:
   (a) Tirer un point de vue $v_d \sim \text{Discrète}(\boldsymbol{\pi})$;

(b) Pour chaque phrase $m = 1, ..., M_d$:

i. Tirer un thème $z_{dm} \sim \text{Discrète}(\boldsymbol{\theta}_{v_d})$;

ii. Pour chaque emplacement de mots $n = 1, ..., N_{dm}$:

A. Choisir une catégorie de partie du discours x_{dmn} [3] parmi $\{0, 1\}$;

B. Si $x_{dmn} = 0$, tirer un mot thématique $w_{dmn} \sim \text{Discrète}(\boldsymbol{\phi}_{0z_{dm}})$,
Sinon si $x_{dmn} = 1$, tirer un mot d'opinion $w_{dmn} \sim \text{Discrète}(\boldsymbol{\phi}_{1v_dz_{dm}})$.

La probabilité jointe des variables aléatoires du modèle $(\mathbf{w}, \mathbf{z}, \mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \boldsymbol{\pi})$ s'obtient de manière immédiate en utilisant les dépendances statistiques définies dans le modèle graphique (Figure 4.1) et l'histoire générative de VODUM :

$$
\begin{aligned}
& p(\mathbf{w}, \mathbf{z}, \mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\phi}_0, \boldsymbol{\phi}_1, \boldsymbol{\pi} \,; \mathbf{x}, \alpha, \beta_0, \beta_1, \gamma) \qquad (4.1) \\
= & \left\{ \prod_{d=1}^{D} \pi_{v_d} \prod_{m=1}^{M_d} \theta_{v_d z_{dm}} \prod_{n=1}^{N_{dm}} [\phi_{0 z_{dm} w_{dmn}}]^{\mathbb{I}(x_{dmn}=0)} [\phi_{1 v_d z_{dm} w_{dmn}}]^{\mathbb{I}(x_{dmn}=1)} \right\} \\
& \left\{ \prod_{v=1}^{V} \text{Dir}_T(\boldsymbol{\theta}_v \,; \alpha) \right\} \left\{ \prod_{z=1}^{T} \text{Dir}_W(\boldsymbol{\phi}_{0z} \,; \beta_0) \right\} \left\{ \prod_{v=1}^{V} \prod_{z=1}^{T} \text{Dir}_W(\boldsymbol{\phi}_{1vz} \,; \beta_1) \right\} \text{Dir}_V(\boldsymbol{\pi} \,; \eta).
\end{aligned}
$$

où Dir dénote une loi de Dirichlet et $\mathbb{I}(\cdot)$ désigne la fonction indicatrice : $\mathbb{I}(\text{vrai}) = 1$ et $\mathbb{I}(\text{faux}) = 0$.

Similairement aux modèlés thématiques probabilistes dérivés de LDA, l'inférence postérieure exacte de VODUM n'est pas réalisable (cf. 3.2.3). Par conséquent, nous faisons appel à une technique d'inférence approchée pour estimer les distributions $\boldsymbol{\phi}_0$, $\boldsymbol{\phi}_1$, $\boldsymbol{\theta}$ et $\boldsymbol{\pi}$, ainsi que les attributions de points de vue aux documents $\mathbf{v}$ et de thèmes aux phrases $\mathbf{z}$. Nous détaillons cette procédure dans la Section 4.1.2.

4.1.2 Inférence postérieure

Afin de réaliser l'inférence postérieure approchée de notre modèle VODUM, nous avons choisi d'utiliser un échantillonneur de Gibbs marginalisé [Liu, 1994] pour sa simplicité d'implémentation. L'échantillonneur de Gibbs marginalisé, que nous avons décrit en détail dans la Section 3.3.1, est un algorithme de Monte Carlo par chaînes de Markov permettant de générer un ensemble de tirages issues de la loi postérieure d'un modèle après marginalisation d'un ensemble de variables latentes. Dans le cas du modèle VODUM, la loi postérieure après marginalisation des distributions $\boldsymbol{\phi}_0$, $\boldsymbol{\phi}_1$, $\boldsymbol{\theta}$ et $\boldsymbol{\pi}$ est $p(\mathbf{v}, \mathbf{z} | \mathbf{w}, \mathbf{x} \,; \alpha, \beta_0, \beta_1, \eta)$, c'est-à-dire la loi jointe des variables latentes $\mathbf{v}$ (attributions de points de vue aux documents) et $\mathbf{z}$ (attributions de thèmes aux phrases) conditionnée sur les variables observées $\mathbf{w}$ (mots de la collection) et $\mathbf{x}$ (catégories de partie du discours des mots). Pour estimer la loi postérieure de VODUM, la mise en œuvre de l'échantillonneur de Gibbs marginalisé requiert seulement de pouvoir

3. Les variables $\mathbf{x}$ sont ici traitées comme des variables déterministes et non comme des variables aléatoires. Ce choix s'explique par le fait que la distribution observée des catégories de parties de discours ne constitue pas une information utile à notre sujet d'étude : les points de vue.

calculer la probabilité de chaque variable latente conditionnée sur les autres variables latentes et sur les variables observées. L'échantillonneur de Gibbs marginalisé consiste alors à effectuer successivement des tirages de chaque variable latente suivant cette loi conditionnelle pour un nombre d'itérations donné.

Ainsi, la clé de l'échantillonneur de Gibbs marginalisé de VODUM est le calcul des probabilités $p(z_{dm}|\mathbf{v}, \mathbf{z}^{-(dm)}, \mathbf{w}, \mathbf{x}\,; \alpha, \beta_0, \beta_1, \eta)$ (pour chaque document $d = 1, ..., D$ et phrase $m = 1, ..., M_d$) et $p(v_d|\mathbf{v}^{-d}, \mathbf{z}, \mathbf{w}, \mathbf{x}\,; \alpha, \beta_0, \beta_1, \eta)$ (pour chaque document $d = 1, ..., D$), que nous développons dans les Équations (4.2) et (4.4), respectivement. Les notations utilisées dans les équations qui suivent sont définies dans le Tableau 4.1. En outre, les exposants $-d$ et $-(dm)$ excluent des compteurs ou des variables le document d et la phrase m du document d, respectivement. Similairement, les exposants (d) et (dm) indiquent respectivement que seuls le document d et la phrase m du document d sont pris en compte. Enfin, le symbole $\bullet$ en indice dans un compteur indique une sommation pour la variable à la place correspondante (par exemple, $n_{v\bullet} = \sum_{z=1}^{T} n_{vz}$).

Calcul de $p(z_{dm}|\mathbf{v}, \mathbf{z}^{-(dm)}, \mathbf{w}, \mathbf{x}\,; \alpha, \beta_0, \beta_1, \eta)$. La probabilité conditionnelle des attributions de thèmes aux phrases peut être calculée de la manière suivante pour $d = 1, ..., D$, $m = 1, ..., M_d$, $z = 1, ..., T$:

$$
\begin{aligned}
& p(z_{dm} = z|v_d = v, \mathbf{v}^{-(d)}, \mathbf{z}^{-(dm)}, \mathbf{w}, \mathbf{x}\,; \alpha, \beta_0, \beta_1, \eta) \qquad (4.2)\\
\propto\ & \frac{p(\mathbf{z}|\mathbf{v}\,; \alpha)}{p(\mathbf{z}^{-(dm)}|\mathbf{v}\,; \alpha)} \cdot \frac{p(\mathbf{w}|\mathbf{v}, \mathbf{z}, \mathbf{x}\,; \beta_0, \beta_1)}{p(\mathbf{w}^{-(dm)}|\mathbf{v}, \mathbf{z}^{-(dm)}, \mathbf{x}^{-(dm)}\,; \beta_0, \beta_1)}\\
=\ & \frac{\Gamma(n_{v\bullet}^{-(dm)} + T\alpha)}{\Gamma(n_{v\bullet} + T\alpha)} \cdot \frac{\Gamma(n_{vz} + \alpha)}{\Gamma(n_{vz}^{-(dm)} + \alpha)} \cdot \frac{\Gamma(n_{0z\bullet}^{-(dm)} + W_0\beta_0)}{\Gamma(n_{0z\bullet} + W_0\beta_0)} \cdot \prod_{w \in \mathcal{W}_0} \frac{\Gamma(n_{0zw} + \beta_0)}{\Gamma(n_{0zw}^{-(dm)} + \beta_0)}\\
& \cdot \frac{\Gamma(n_{1vz\bullet}^{-(dm)} + W_1\beta_1)}{\Gamma(n_{1vz\bullet} + W_1\beta_1)} \cdot \prod_{w \in \mathcal{W}_1} \frac{\Gamma(n_{1vzw} + \beta_1)}{\Gamma(n_{1vzw}^{-(dm)} + \beta_1)}\\
=\ & \frac{n_{vz}^{-(dm)} + \alpha}{n_{v\bullet}^{-(dm)} + T\alpha} \cdot \frac{\prod_{w \in \mathcal{W}_0} \prod_{a=0}^{n_{0zw}^{(dm)}-1} (n_{0zw}^{-(dm)} + \beta_0 + a)}{\prod_{b=0}^{n_{0z\bullet}^{(dm)}-1} (n_{0z\bullet}^{-(dm)} + W_0\beta_0 + b)} \cdot \frac{\prod_{w \in \mathcal{W}_1} \prod_{a=0}^{n_{1vzw}^{(dm)}-1} (n_{1vzw}^{-(dm)} + \beta_1 + a)}{\prod_{b=0}^{n_{1vz\bullet}^{(dm)}-1} (n_{1vz\bullet}^{-(dm)} + W_1\beta_1 + b)}
\end{aligned}
$$

La dérivation présentée dans l'Équation (4.2) est similaire à celle effectuée dans le cadre de l'échantillonneur de Gibbs marginalisé de LDA, décrite dans l'Équation (3.12) de la Section 3.3.1.3. On retrouve dans l'expression de la dernière ligne trois termes :

— Le premier terme reflète la fréquence de l'attribution du thème z parmi les phrases des documents de même point de vue v que le document courant.

— Les deuxième et troisième termes expriment dans quelle mesure les mots respectivement thématiques et d'opinion de la phrase courante apparaissent également dans des phrases assignées à z dans le reste de la collection.

On note toutefois que les deuxième et troisième termes sont plus complexes qu'un simple ratio de compteurs lissés par le *prior* de Dirichlet (tel que le premier terme). Cela est dû au fait que la différence entre les compteurs de mots incluant et excluant la phrase courante peut être supérieure à 1 (une phrase pouvant contenir plusieurs mots). Par exemple, $n_{0zw} - n_{0zw}^{-(dm)} = n_{0zw}^{(dm)}$ correspond au nombre d'occurrences du mot thématique w dans la phrase m du document d, qui est potentiellement plus grand que 1. Par conséquent, la propriété $\Gamma(x+1) = x\,\Gamma(x)$ pour $x \in \mathbb{R}$ n'est pas utilisable dans ce cas ; il faut alors faire appel à l'identité plus générale $\Gamma(x+n) = \Gamma(x)\prod_{i=0}^{n-1}(x+i)$ pour $x \in \mathbb{R}$ et $n \in \mathbb{N}$. Appliquée à l'exemple précédent, cela donne alors le produit suivant [4] :

$$\frac{\Gamma(n_{0zw}+\beta_0)}{\Gamma(n_{0zw}^{-(dm)}+\beta_0)} = \frac{\Gamma(n_{0zw}^{-(dm)}+n_{0zw}^{(dm)}+\beta_0)}{\Gamma(n_{0zw}^{-(dm)}+\beta_0)} = \prod_{a=0}^{n_{0zw}^{(dm)}-1}(n_{0zw}^{-(dm)}+\beta_0+a). \tag{4.3}$$

Calcul de $p(v_d|\mathbf{v}^{-(d)},\mathbf{z},\mathbf{w},\mathbf{x}\,;\alpha,\beta_0,\beta_1,\eta)$. La probabilité conditionnelle des attributions de points de vue aux documents est calculée de manière similaire, avec $d = 1,\ldots,D$, $v = 1,\ldots,V$:

$$\begin{aligned}
&p(v_d = v|\mathbf{v}^{-(d)},\mathbf{z},\mathbf{w},\mathbf{x}\,;\alpha,\beta_0,\beta_1,\eta) \qquad (4.4)\\
&\propto \frac{p(\mathbf{v}\,;\eta)}{p(\mathbf{v}^{-(d)}\,;\eta)}\cdot\frac{p(\mathbf{z}|\mathbf{v}\,;\alpha)}{p(\mathbf{z}^{-(d)}|\mathbf{v}^{-(d)}\,;\alpha)}\cdot\frac{p(\mathbf{w}|\mathbf{v},\mathbf{z},\mathbf{x}\,;\beta_0,\beta_1)}{p(\mathbf{w}^{-(d)}|\mathbf{v}^{-(d)},\mathbf{z}^{-(d)},\mathbf{x}^{-(d)}\,;\beta_0,\beta_1)}\\
&\propto \frac{\Gamma(n_{\bullet}^{-(d)}+V\eta)}{\Gamma(n_{\bullet}+V\eta)}\cdot\frac{\Gamma(n_v+\eta)}{\Gamma(n_v^{-(d)}+\eta)}\cdot\frac{\Gamma(n_{v\bullet}^{-(d)}+T\alpha)}{\Gamma(n_{v\bullet}+T\alpha)}\cdot\prod_{z=1}^{T}\frac{\Gamma(n_{vz}+\alpha)}{\Gamma(n_{vz}^{-(d)}+\alpha)}\\
&\quad\cdot\prod_{z=1}^{T}\left\{\frac{\Gamma(n_{1vz\bullet}^{-(dm)}+W_1\beta_1)}{\Gamma(n_{1vz\bullet}+W_1\beta_1)}\cdot\prod_{w\in\mathcal{W}_1}\frac{\Gamma(n_{1vzw}+\beta_1)}{\Gamma(n_{1vzw}^{-(dm)}+\beta_1)}\right\}\\
&= \frac{n_v^{-(d)}+\eta}{n_{\bullet}^{-(d)}+V\eta}\cdot\frac{\prod_{z=1}^{T}\prod_{a=0}^{n_{vz}^{(d)}-1}(n_{vz}^{-(d)}+\alpha+a)}{\prod_{b=0}^{n_{v\bullet}^{(d)}-1}(n_{v\bullet}^{-(d)}+T\alpha+b)}\cdot\prod_{z=1}^{T}\frac{\prod_{w\in\mathcal{W}_1}\prod_{a=0}^{n_{1zvw}^{(d)}-1}(n_{1vzw}^{-(d)}+\beta_1+a)}{\prod_{b=0}^{n_{1vz\bullet}^{(d)}-1}(n_{1vz\bullet}^{-(d)}+W_1\beta_1+b)}
\end{aligned}$$

On distingue de nouveau trois termes dans l'expression finale de la probabilité conditionnelle des points de vue :

— Le premier terme reflète la fréquence de l'attribution du point de vue v parmi les documents de la collection autre que le document courant.

— Le deuxième terme dénote la propension des documents de point de vue v à contenir des phrases attribués aux mêmes thèmes que les phrases du document courant.

— Le troisième terme exprime dans quelle mesure les mots d'opinion du document courant apparaissent également dans des documents assignés à v dans le reste de la collection.

4. Du point de vue de l'implémentation, seul le logarithme de ce type de produits est stocké en mémoire afin d'éviter les problèmes de dépassement numérique.

Estimation de $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$. À partir des tirages successivement collectés suivant les Équations (4.2) et (4.4), il est possible de construire des estimateurs bayésiens pour les distributions $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$ comme suit :

$$\hat{\pi}_v = E\left[\pi_v | \mathbf{v}\,; \eta\right] = \frac{n_v + \eta}{n_\bullet + V\eta}\,; \tag{4.5}$$

$$\hat{\theta}_{vz} = E\left[\theta_{vz} | \mathbf{v}, \mathbf{z}\,; \alpha\right] = \frac{n_{vz} + \alpha}{n_{v\bullet} + T\alpha}\,; \tag{4.6}$$

$$\hat{\phi}_{0zw} = E\left[\phi_{0zw} | \mathbf{z}, \mathbf{w}, \mathbf{x}\,; \beta_0\right] = \begin{cases} \dfrac{n_{0zw} + \beta_0}{n_{0z\bullet} + W_0\beta_0} & \text{si } w \in \mathcal{W}_0, \\ 0 & \text{sinon ;} \end{cases} \tag{4.7}$$

$$\hat{\phi}_{1vzw} = E\left[\phi_{1vzw} | \mathbf{v}, \mathbf{z}, \mathbf{w}, \mathbf{x}\,; \beta_1\right] = \begin{cases} \dfrac{n_{1vzw} + \beta_1}{n_{1vz\bullet} + W_1\beta_1} & \text{si } w \in \mathcal{W}_1, \\ 0 & \text{sinon ;} \end{cases} \tag{4.8}$$

avec $v = 1, ..., V$, $z = 1, ..., T$, $w = 1, ..., W$.

Résumé de l'algorithme complet. Nous pouvons maintenant fournir une description complète de l'échantillonneur de Gibbs marginalisé pour VODUM, résumée dans l'Algorithme 4 [5]. Tout d'abord, les points de vue latents $\mathbf{v}$ de chaque document et les thèmes latents $\mathbf{z}$ de chaque phrase sont initialisés en effectuant des tirages suivant les lois discrètes uniformes Uniforme$(1, ..., V)$ et Uniforme$(1, ..., T)$, respectivement. Ensuite, la chaîne de Markov dont la distribution stationnaire est la loi postérieure de VODUM est construite itérativement comme suit. Pour chaque itération $s = 1, ..., S$, la collection est parcourue en tirant pour chaque document d le point de vue $v_d^{(s)}$ suivant l'Équation (4.4), et pour chaque phrase m de d le thème $z_{dm}^{(s)}$ suivant l'Équation (4.2). Enfin, après parcours de la collection, les différentes distributions sont estimées en calculant $\hat{\pi}_v^{(s)}$, $\hat{\theta}_{vz}^{(s)}$, $\hat{\phi}_{0zw}^{(s)}$ et $\hat{\phi}_{1vzw}^{(s)}$ à partir des Équations (4.5), (4.6), (4.7) et (4.8). Après les S itérations, l'algorithme retourne les différentes tirages et estimateurs $\left\{\mathbf{v}^{(s)}\right\}_{s=1}^{S}$, $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\pi}}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\theta}}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\phi}}_0^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\phi}}_1^{(s)}\right\}_{s=1}^{S}$.

À partir du modèle VODUM appris suivant la procédure décrite dans cette section, nous présentons son évaluation et sa comparaison aux modèles de l'état de l'art dans la Section 4.2.

4.2 Expérimentations

À travers les expérimentations que nous avons conduites sur notre modèle VODUM, nous avons étudié les quatre hypothèses suivantes postulant que la capacité de VODUM à identifier les points de vue est positivement impactée par :

5. Pour augmenter la lisibilité de l'algorithme, nous n'explicitons pas la mise à jour des différents compteurs (tels que n_{vz}). Les opérations qu'il faudrait appliquer à ces compteurs sont similaires à celles décrites dans l'Algorithme 3.

Algorithme 4 : Échantillonneur de Gibbs marginalisé pour VODUM

Input : un nombre d'itérations S, les hyperparamètres α, β_0, β_1 et η, le nombre de thèmes T, le nombre de points de vue V, les mots de la collection $\mathbf{w}$, les catégories de partie du discours des mots $\mathbf{x}$

Output : S tirages approximativement distribués selon $p(\mathbf{v}, \mathbf{z}|\mathbf{w}, \mathbf{x}\,; \alpha, \beta_0, \beta_1, \eta)$, S estimations des distributions $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$

// Initialisation aléatoire de $\mathbf{v}^{(0)}$ et $\mathbf{z}^{(0)}$

for $d \leftarrow 1$ **to** D **do**
  Tirer $v_d^{(0)} \sim \text{Uniforme}(1, ..., V)$
  for $m \leftarrow 1$ **to** M_d **do**
    Tirer $z_{dn}^{(0)} \sim \text{Uniforme}(1, ..., T)$
  end
end

// Construction de la chaîne de Markov

for $s \leftarrow 1$ **to** S **do**
  // Tirage de $\mathbf{v}^{(s)}$ et $\mathbf{z}^{(s)}$
  for $d \leftarrow 1$ **to** D **do**
    Tirer $v_d^{(s)}$ selon l'Équation (4.4)
    for $m \leftarrow 1$ **to** M_d **do**
      Tirer $z_{dm}^{(s)}$ selon l'Équation (4.2)
    end
  end
  // Estimation de $\boldsymbol{\pi}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$
  for $v \leftarrow 1$ **to** V **do**
    Calculer $\hat{\pi}_v^{(s)}$ selon l'Équation (4.5)
  end
  for $v \leftarrow 1$ **to** V, $z \leftarrow 1$ **to** T **do**
    Calculer $\hat{\theta}_{vz}^{(s)}$ selon l'Équation (4.6)
  end
  for $z \leftarrow 1$ **to** T, $w \leftarrow 1$ **to** W **do**
    Calculer $\hat{\phi}_{0zw}^{(s)}$ selon l'Équation (4.7)
  end
  for $v \leftarrow 1$ **to** V, $z \leftarrow 1$ **to** T, $w \leftarrow 1$ **to** W **do**
    Calculer $\hat{\phi}_{1vzw}^{(s)}$ selon l'Équation (4.8)
  end
end

return $\left\{\mathbf{v}^{(s)}\right\}_{s=1}^{S}$, $\left\{\mathbf{z}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\pi}}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\theta}}^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\phi}}_0^{(s)}\right\}_{s=1}^{S}$, $\left\{\hat{\boldsymbol{\phi}}_1^{(s)}\right\}_{s=1}^{S}$

— **(H1)** l'utilisation de distributions de thèmes spécifiques aux points de vue (au lieu d'être spécifiques aux documents comme dans LDA) ;

— **(H2)** la séparation entre les mots d'opinion et les mots thématiques ;

— **(H3)** l'utilisation de variables de thème au niveau de la phrase ;

— **(H4)** l'utilisation de variables de point de vue au niveau du document.

En outre, nous testons l'hypothèse **(H5)** selon laquelle VODUM obtient de meilleures performances dans les tâches de modélisation et de regroupement de points de vue que les modèles de l'état de l'art (que nous détaillerons dans la Section 4.2.1.1).

Notons qu'un problème similaire à (H1) avait été abordé dans [Qiu et Jiang, 2013; Qiu *et al.*, 2013b]. Les auteurs n'ont cependant pas évalué l'impact de cette hypothèse sur une tâche de regroupement de points de vue – ce que nous faisons ici. Le reste de cette section est organisé comme suit. Nous décrivons tout d'abord dans la Section 4.2.1 le cadre expérimental que nous avons adopté. Puis, dans la Section 4.2.2, nous détaillons le résultat de l'évaluation des différents modèles.

4.2.1 Cadre expérimental

Dans cette section, nous détaillons le cadre de nos expérimentations, en explicitant les modèles auxquels nous nous comparons (Section 4.2.1.1), la collection de données utilisée (Section 4.2.1.2) et le choix des paramètres effectué (Section 4.2.1.3)

4.2.1.1 Modèles de référence

Nous avons comparé VODUM à plusieurs modèles de l'état de l'art non supervisés ainsi qu'à différentes versions dégénérées de notre modèle afin de répondre aux questions de recherche sous-jacentes à nos cinq hypothèses. Les modèles de référence que nous avons considéré pour étudier (H5) sont LDA [Blei *et al.*, 2001], JTV [Trabelsi et Zaïane, 2014] et TAM [Paul et Girju, 2010; Paul *et al.*, 2010].

— **LDA** [Blei *et al.*, 2001, 2003] : l'allocation de Dirichlet latente (*latent Dirichlet allocation* – LDA) est le modèle thématique probabiliste de base que nous avons amplement décrit dans le Chapitre 3. Nous choisissons de comparer VODUM à ce modèle simple – qui a été défini pour modéliser les thèmes et non les points de vue – afin d'étudier si la découverte de points de vue peut être réalisée par les mêmes techniques que la découverte de thèmes.

— **JTV** [Trabelsi et Zaïane, 2014] : le modèle *joint topic viewpoint* (JTV), introduit dans la Section 2.2.1, modélise conjointement les thèmes et les points de vue et a été initialement développer pour la détection d'expressions de contention. À la différence de VODUM, JTV définit cependant les attributions de thèmes et de points de vue au niveau des mots et les distributions de points de vue au niveau des documents. De plus, dans JTV, les attributions de points de vue sont conditionnées par les attributions de thèmes (et

non l'inverse comme dans VODUM). Enfin, JTV considère que tous les mots sont des mots d'opinion, sans distinguer mots d'opinion et mots thématiques.

— **TAM** [Paul et Girju, 2010] : le *topic-aspect model* (TAM), présenté dans la Section 2.3.1.1, a été initialement développé pour modéliser les thèmes et les aspects – qui représentent les points de vue dans notre contexte. Les attributions de thèmes et d'aspects sont positionnés au niveau des mots dans TAM. Les mots peuvent être tirés suivant quatre distributions différentes : distribution de fond (indépendante des thèmes et des points de vue), distributions de mots thématiques (dépendantes des thèmes), distributions de mots d'opinion (dépendantes des points de vue), distributions de mots thématiques d'opinion (dépendantes des thèmes et des points de vue). Cependant, à la différence de VODUM, le fait qu'un mot provienne d'un type de distributions ou d'un autre est basé sur des variables latentes (aléatoire) plutôt que sur les catégories de parties du discours (déterministes).

Les quatre versions dégénérées de VODUM ont quant à elles été définies pour évaluer l'impact de chacune des composantes de notre modèle de manière isolée. Nous illustrons la représentation sous forme de modèle graphique des différentes versions dégénérées dans la Figure 4.2 et leur but est détaillé ci-dessous :

— **VODUM-D** (Figure 4.2a) : dans VODUM-D, les distributions de thèmes ont été définies au niveau du document. Dans VODUM, les distributions sont au contraire spécifiques au document et indépendante des documents. VODUM-D a été défini pour étudier (H1).

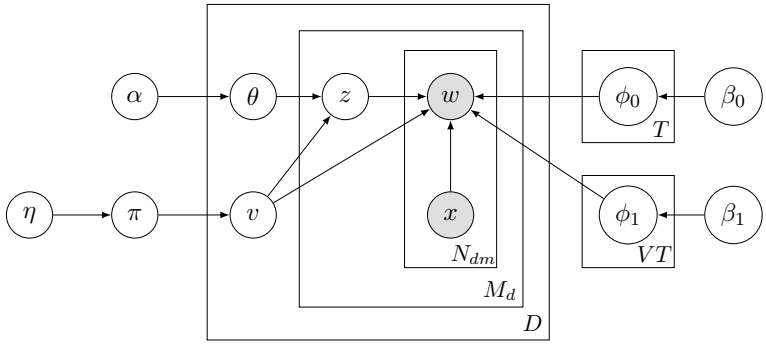
— **VODUM-O** (Figure 4.2b) : VODUM-O suppose que tous les mots sont des mots d'opinion, c'est-à-dire, que tous les mots sont tirés de distributions dépendant à la fois du point de vue et du thème. Au contraire, VODUM fait la distinction entre les mots d'opinion (tirés de distributions spécifiques aux points de vue et aux thèmes) et les mots thématiques (tirés de distributions dépendant uniquement des thèmes). La comparaison de VODUM et de VODUM-O permet de répondre à (H2).

— **VODUM-W** (Figure 4.2c) : VODUM-W définit les attributions de thèmes au niveau du mot, au lieu du niveau de la phrase comme dans VODUM. Cela permet au document d'être potentiellement associé à plus de thèmes différents (un thème par mot au lieu d'un thème par phrase). VODUM-W est défini pour traiter (H3).

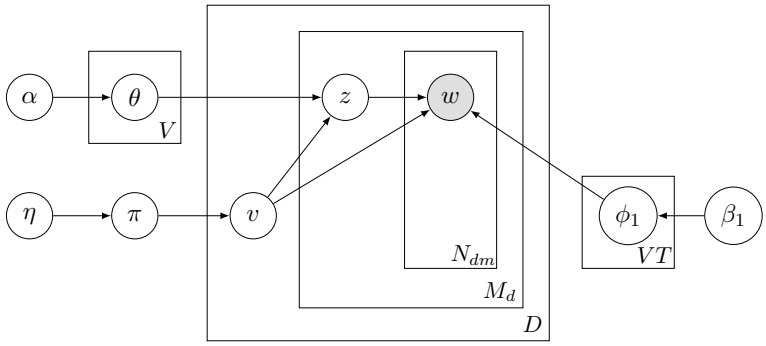
— **VODUM-S** (Figure 4.2d) : VODUM-S modélise les attributions de points de vue au niveau de la phrase alors que dans VODUM les attributions de points de vue sont définis au niveau du document. Par conséquent, un document modélisé par VODUM-S peut contenir des phrases attribuées à différents points de vue. Comparer VODUM et VODUM-S permet d'apporter une réponse à (H4).

Nous avons implémenté VODUM et les différents modèles de référence dans le langage Java en nous basant sur JGibbLDA [6], qui propose un échantillonneur de Gibbs marginalisé pour LDA. Le code source de notre implémentation et les données formatées (après les différentes étapes de prétraitement, décrites en Section 4.2.1.2) sont publiquement disponibles sur `https://github.com/tthonet/VODUM`.

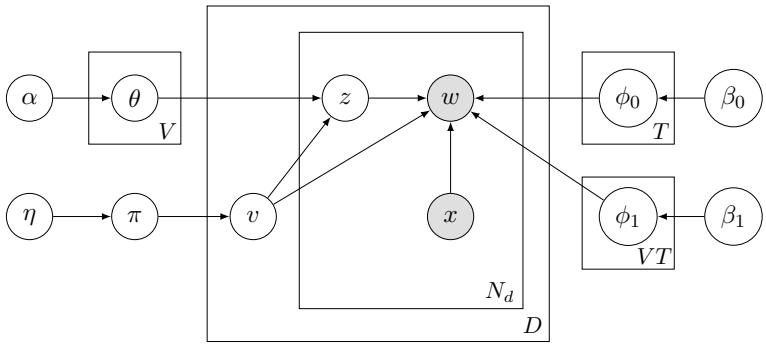
6. `http://jgibblda.sourceforge.net/`



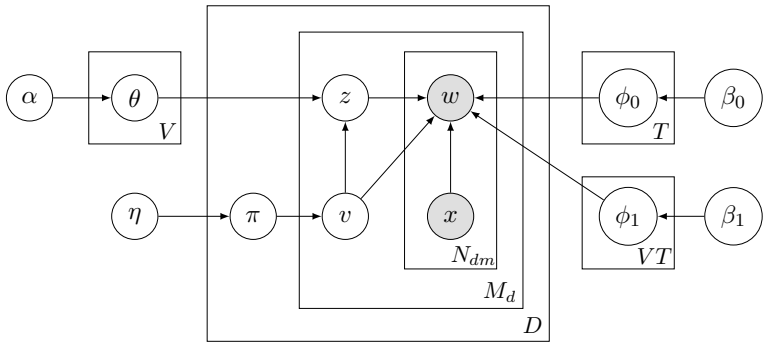
(a) VODUM-D : distributions de thèmes au niveau des documents.



(b) VODUM-O : tous les mots sont des mots d'opinions.



(c) VODUM-W : attributions de thèmes au niveau des mots.



(d) VODUM-S : attributions de points de vue au niveau des phrases.

FIGURE 4.2 – Modèles graphiques des versions dégénérées de VODUM.

4.2.1.2 Collection de données

Nous avons comparé VODUM aux modèles de référence en nous appuyant sur la collection Bitterlemons, que nous avons décrite dans la Section 2.3.1.1. La collection Bitterlemons contient des essais en langue anglaise publiés dans l'*e-zine* éponyme [7] et rédigés par des auteurs israéliens et palestiniens, discutant du conflit israélo-palestinien et de sujets connexes. Elle a été introduite par [Lin *et al.*, 2006] puis utilisée par la suite dans de nombreux travaux dont l'objectif est d'identifier et de modéliser les points de vue dans le texte (par exemple, [Paul et Girju, 2010; Trabelsi et Zaïane, 2014]). Bitterlemons contient 297 essais écrits par des auteurs israéliens et 297 écrits par des auteurs palestiniens. Avant d'expérimenter sur cette collection, nous effectuons les prétraitements suivants en utilisant la bibliothèque Lingpipe [8]. Tout d'abord, nous éliminons les caractères numériques. Nous retirons ensuite les mots vides et appliquons la racinisation (*stemming*) de Porter. En faisant appel à un étiqueteur de parties du discours, nous annotons ensuite chaque mot de la collection avec sa catégorie de parties du discours (0 ou 1) comme décrit dans la Section 4.1.1. La catégorie 0 correspond aux mots thématiques et contient les noms communs et les noms propres. La catégorie 1 correspond aux mots d'opinions et contient les adjectifs, verbes, adverbes, modaux et prépositions. Les mots étiquetés avec une partie du discours autre que celles-ci sont supprimés de la collection.

4.2.1.3 Choix des paramètres

Nous décrivons dans cette section les valeurs des paramètres utilisées pour VODUM et les modèles de référence. Les hyperparamètres de VODUM ont été fixés à l'issue d'expérimentations préliminaires : $\alpha = 0.01$, $\beta_0 = \beta_1 = 0.01$ et $\eta = 100$. Le choix d'une faible valeur pour α (l'hyperparamètre lié à $\boldsymbol{\theta}_v$) et d'une valeur élevée pour η (l'hyperparamètre lié à $\boldsymbol{\pi}$) est motivé par le fait que l'on souhaite avoir des distributions $\{\boldsymbol{\theta}_v\}_{v=1}^{V}$ parcimonieuses (*sparse*) – chaque point de vue est associé à un nombre limité de thèmes – et une distribution $\boldsymbol{\pi}$ plus lisse – chaque document a une chance égale d'être généré suivant chacun des points de vue. Nous avons adopté les mêmes hyperparamètres pour les versions dégénérées de VODUM (VODUM-D, VODUM-O, VODUM-W, VODUM-S). Les hyperparamètres de TAM ont été fixés suivant [Paul *et al.*, 2010] : $\alpha = 0.1$, $\beta = 0.1$, $\delta_0 = 80.0$, $\delta_1 = 20.0$, $\gamma_0 = \gamma_1 = 5.0$, $\omega = 0.01$. Pour JTV, nous avons utilisé les valeurs des hyperparamètres fournies dans [Trabelsi et Zaïane, 2014] : $\alpha = 0.01$, $\beta = 0.01$, $\gamma = 25$. Les hyperparamètres de LDA ont été définis comme suit : $\alpha = 0.5$ et $\beta = 0.01$. Pour l'ensemble des expérimentations, nous avons fixé le nombre de points de vue (pour VODUM, VODUM-D, VODUM-O, VODUM-W, VODUM-S et JTV) et le nombre d'aspects (pour TAM) à 2, étant donné que les documents de la collection Bitterlemons sont supposés refléter les points de vue israélien et palestinien.

7. `http://www.bitterlemons.net/`
8. `http://alias-i.com/lingpipe/`

4.2.2 Évaluation

Nous avons conduit une évaluation à la fois quantitative et qualitative pour estimer la qualité de notre modèle. L'évaluation quantitative, comparant VODUM aux différents modèles de référence, repose sur deux métriques : la perplexité (Section 4.2.2.1) et l'exactitude du regroupement de points de vue (Section 4.2.2.2). En outre, l'évaluation qualitative consiste à examiner la cohérence des mots thématiques et des mots d'opinion obtenus pour le même thème par notre modèle VODUM (Section 4.2.2.3). Ces différentes évaluations sont décrites en détail dans le reste de cette section.

4.2.2.1 Perplexité

La perplexité, que nous avons précédemment décrit dans la Section 3.4.1, est une métrique fréquemment utilisée pour mesurer la capacité de généralisation d'un modèle thématique [Blei *et al.*, 2001, 2003]. La perplexité peut être interprétée comme l'inverse de la moyenne géométrique de la log-vraisemblance. Similairement à LDA, la perplexité de VODUM peut être calculée en se basant sur les estimateurs décrits dans les Équations (4.5), (4.6), (4.7) et (4.8) [9]. La méthode que nous avons utilisée pour le calcul de la perplexité est celle de Heinrich [2008] : le modèle est entraîné sur un ensemble d'apprentissage, puis, en gardant le modèle ainsi appris (c'est-à-dire sans le réinitialiser), l'inférence est « continuée » sur les documents de l'ensemble de test (en remplaçant l'ensemble d'apprentissage par celui-ci). Une perplexité plus basse pour l'ensemble de test indique une meilleure capacité du modèle à généraliser à partir des données d'apprentissage. La perplexité de VODUM se calcule de la manière suivante :

$$\text{perp}(\mathbf{w}^{\text{test}}) = \exp\left\{-\frac{\sum_{d=1}^{D^{\text{test}}} \log \hat{p}(w_d)}{\sum_{d=1}^{D^{\text{test}}} \sum_{m=1}^{M_d} N_{dm}}\right\} \tag{4.9}$$

avec :

$$\hat{p}(w_d) = \sum_{v=1}^{V} \hat{\pi}_v \prod_{m=1}^{M_d} \left\{ \sum_{z=1}^{T} \hat{\theta}_{vz} \prod_{n=1}^{N_{dm}} \left[\hat{\phi}_{0zw_{dmn}}\right]^{\mathbb{I}(x_{dmn}=0)} \left[\hat{\phi}_{1vzw_{dmn}}\right]^{\mathbb{I}(x_{dmn}=1)} \right\}. \tag{4.10}$$

Dans cette expérimentation, nous avons pour objectif d'étudier (H5) et de comparer la capacité de généralisation de notre modèle VODUM à celle des modèles de référence. Nous avons pour cela utilisé une validation croisée à 10 plis comme suit. Chaque modèle a été entraîné sur 9 plis de la collection Bitterlemons (ensemble d'apprentissage) pendant 1000 itérations, puis l'inférence a été poursuivie sur le pli restant (ensemble de test) pour 1000 itérations supplémentaires. Nous répétons cette procédure 10 fois en changeant le pli de test – pour faire en sorte que chaque pli ait été utilisé comme pli de test – et nous reportons la perplexité moyennée sur les 10 plis de test. Étant donné que la performance de généralisation dépend du nombre de thèmes – qui influence le nombre de paramètres et donc la capacité

9. Nous utilisons ici les estimateurs basés sur les tirages de la dernière itération S.

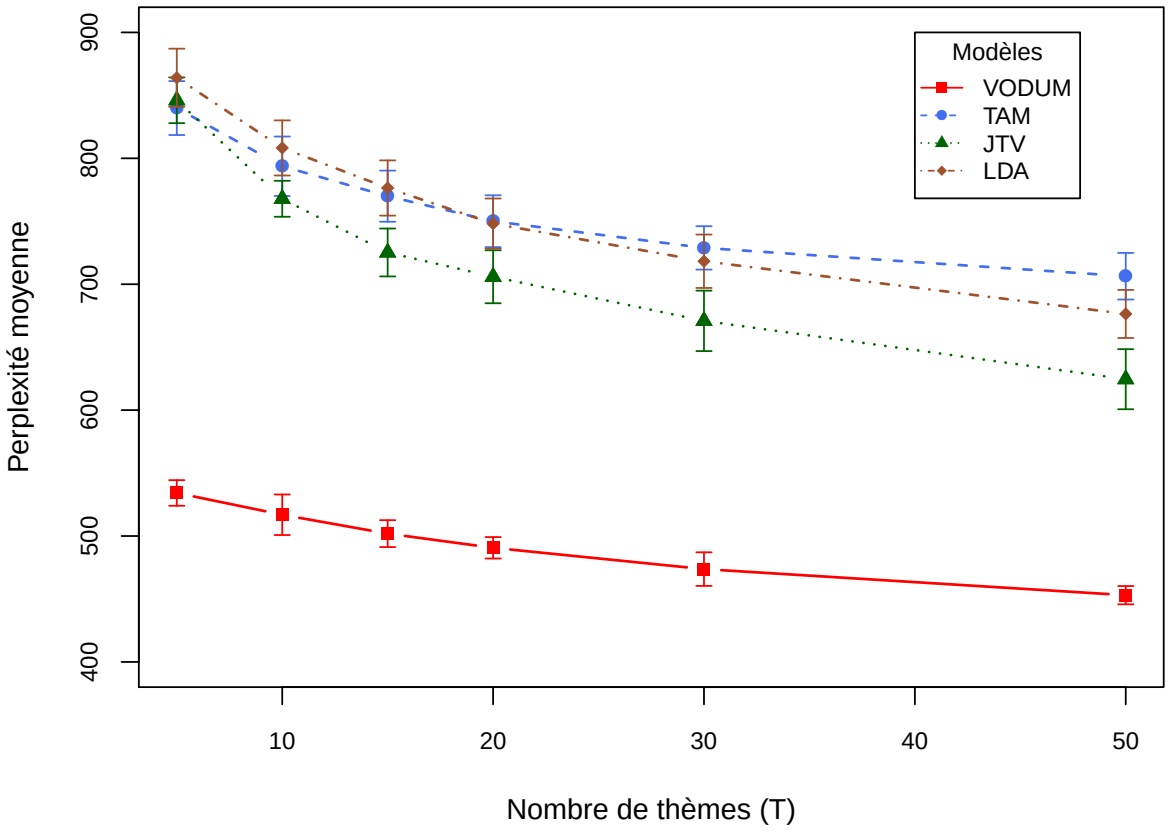


FIGURE 4.3 – Perplexité des modèles VODUM, TAM, JTV et LDA calculée pour 5, 10, 15, 20, 30 et 50 thèmes. Une valeur plus basse indique une meilleure performance. Les barres d'erreur dénotent les intervalles de confiance à 95 % autour de la moyenne, calculés à partir d'une loi t de Student sur les 10 exécutions de la validation croisée.

d'apprentissage des modèles – nous mesurons la perplexité pour différents nombres de thèmes $T = 5, 10, 15, 20, 30, 50$.

Les résultats de cette expérimentation sont illustrés dans la Figure 4.3. Les valeurs numériques sont données en complément dans le Tableau 4.2. Les barres d'erreur indiquent l'intervalle de confiance à 95 % autour de la perplexité moyenne calculée à partir d'une loi t de Student sur les 10 exécutions de la validation croisée. Ainsi, si les intervalles de confiance de deux modèles ne se superposent pas, cela signifie que la différence est statistiquement significative avec un niveau de significativité $\alpha = 0,05$. Pour les différents nombre de thèmes, VODUM a une perplexité significativement plus faible que celle de TAM, JTV et LDA. JTV présente une perplexité légérement plus faible que TAM et LDA, en particulier pour un nombre de thèmes élevé. Les perplexités obtenues par TAM et LDA sont comparables, TAM étant légèrement meilleur (sa perplexité étant plus basse) pour un petit nombre de thèmes et LDA pour un grand nombre de thèmes.

Tableau 4.2 – Perplexité des modèles VODUM, TAM, JTV et LDA calculée pour 5, 10, 15, 20, 30 et 50 thèmes. Une valeur plus basse indique une meilleure performance. Le meilleur résultat est signalé en gras. Les valeurs indiquées après le symbole ± fournissent les intervalles de confiance à 95 % autour de la moyenne, calculés à partir d'une loi t de Student sur les 10 exécutions de la validation croisée.

Modèle	Nombre de thèmes (T)					
	5	10	15	20	30	50
LDA	864 ± 23	808 ± 22	776 ± 22	748 ± 20	718 ± 21	676 ± 19
JTV	846 ± 18	768 ± 14	725 ± 19	706 ± 21	671 ± 24	625 ± 24
TAM	840 ± 21	794 ± 24	770 ± 20	750 ± 21	729 ± 17	706 ± 19
VODUM	$\mathbf{534 \pm 10}$	$\mathbf{517 \pm 16}$	$\mathbf{502 \pm 11}$	$\mathbf{491 \pm 8}$	$\mathbf{474 \pm 13}$	$\mathbf{453 \pm 7}$

Nous suspectons toutefois la perplexité de VODUM d'être plus basse en raison des variables déterministes intégrées à ce modèle – les catégories de partie du discours de chaque mot. En effet, cela a pour effet de séparer le vocabulaire en deux : le vocabulaire des mots thématiques et le vocabulaire des mots d'opinion. Ainsi, la prédictibilité des mots est accrue car leur distribution repose sur un support plus restreint. Cela a pour effet d'augmenter la vraisemblance et de diminuer la perplexité. Par conséquent, la perplexité plus basse de VODUM ne signifie pas nécessairement que le modèle généralise mieux.

Par ailleurs, la perplexité est une mesure statistique qui ne reflète pas nécessairement dans quelle mesure VODUM est capable d'identifier correctement les points de vue attribués à chaque document. Nous comparons donc dans la Section 4.2.2.2 les performances de notre modèle et des modèles de référence dans le cadre d'une tâche de regroupement de points de vue.

4.2.2.2 Regroupement de points de vue

Dans cette expérimentation, nous avons pour objectif d'évaluer l'exactitude d'identification des points de vue (VIA – *viewpoint identification accuracy*) pour VODUM et les modèles de référence afin de tester les hypothèses (H1), (H2), (H3), (H4) et (H5). Étant donné que la collection Bitterlemons contient deux points de vue différents (israélien et palestinien), la mesure de VIA s'apparente ici à l'exactitude d'un problème de regroupement binaire. La VIA est alors la proportion de documents regroupés correctement. Comme il l'a été rapporté dans [Paul *et al.*, 2010], la VIA d'un modèle thématique peut exhiber une variance élevée parmi différentes exécutions de l'échantillonneur de Gibbs, en raison de la nature stochastique de la procédure. Par conséquent, pour chaque modèle évalué, nous effectuons 50 exécutions de l'échantillonneur de Gibbs, avec 1 000 itérations pour chacune. Nous conservons uniquement le tirage final de chaque exécution.

Le nombre de thèmes pour notre modèle VODUM et ses versions dégradées VODUM-D, VODUM-O, VODUM-W et VODUM-S est fixé à 12. Pour les modèles de l'état de l'art, le

nombre de thèmes est choisi en suivant la recommandation de leurs auteurs respectifs : 8 pour TAM (en accord avec [Paul *et al.*, 2010]), 6 pour JTV (en accord avec [Trabelsi et Zaïane, 2014]). Dans le cas de LDA, le nombre de thèmes est fixé à 2 : LDA ne modélisant pas les points de vue, nous étudions dans quelle mesure LDA est capable d'aligner points de vue et thèmes.

Pour VODUM, VODUM-D, VODUM-O et VODUM-W, le dernier tirage de chaque exécution de l'échantillonneur de Gibbs fournit une attribution de points de vue aux documents de la collection. Nous utilisons ainsi directement ces attributions pour former deux groupes représentant chacun un point de vue et évaluer la VIA de ces groupes. VODUM-S a cependant des attributions de points de vue positionnés au niveau des phrases, donc dans ce modèle nous attribuons à chaque document le point de vue qui est majoritaire dans les phrases le composant. Lorsque les phrases d'un document sont assignées de manière équilibrée à chaque point de vue, le point de vue de la phrase est choisi aléatoirement suivant une loi uniforme. Nous avons adopté une approche similaire pour TAM, JTV et LDA en utilisant respectivement la majorité parmi les aspects, points de vue et thèmes (positionnés au niveau des mots) pour déduire les attributions de points de vue des documents.

Le résultat de l'expérimentation est fourni sous forme de boîtes à moustaches dans la Figure 4.4, afin de capturer la variance des résultats. Nous résumons également les VIA moyennes et les intervalles de confiance à 95 % qui leur sont associés (ici calculés à partir des 50 exécutions de l'échantillonneur de Gibbs et suivant une loi t de Student) dans la Table 4.3. On peut observer que le modèle VODUM a globalement obtenu les meilleurs résultats sur la tâche de regroupement de points de vue. Plus spécifiquement, les performances de VODUM sont en moyenne supérieures à celle des modèles de l'état de l'art TAM, JTV et LDA, ce qui soutient l'hypothèse (H5). Notons que VODUM est significativement meilleur (pour un niveau de significativité statistique $\alpha = 0,05$) que JTV et LDA, mais que la différence avec TAM n'est pas significative. Il est probable que cela soit dû à la variance élevée de la VIA, impliquant qu'un nombre supplémentaire d'exécutions pourrait être requis [10].

Parmi les modèles de l'état de l'art, TAM a obtenu les meilleurs résultats. On peut également observer que les performances de JTV ne dépassent pas celle de LDA sur cette tâche. Cela peut sans doute être expliqué par le fait que la dépendance entre attributions de thèmes et de points de vue n'a pas été prise en compte pour identifier les points de vue des documents dans JTV – les attributions de points de vue au niveau des mots dans JTV ne sont pas nécessairement alignées pour des thèmes différents.

Les résultats obtenus par les versions dégénérées soutiennent les hypothèses (H1), (H2), (H3) et (H4). VODUM-O et VODUM-W ont obtenu des performances significativement plus basses que tous les autres modèles, se rapprochant de 50 % (VIA correspondant à une approche qui attribuerait aléatoirement les points de vue aux documents). Par conséquent, la séparation des mots thématiques et des mots d'opinion, ainsi que le positionnement des attributions de thèmes au niveau des phrases – caractéristiques adoptées par VODUM et absentes dans

10. Nous avons ici fixé le nombre d'exécutions à 50 pour limiter le temps de calcul (environ 270 secondes pour une unique exécution).

Tableau 4.3 – Exactitude d'identification des points de vue (VIA) et intervalle de confiance (IC) à 95 % autour de la VIA moyenne pour VODUM, TAM, JTV, LDA, VODUM-D, VODUM-O, VODUM-W et VODUM-S. Une valeur plus élevée indique une meilleure performance. Le meilleur résultat est signalé en gras. Les IC sont calculés à partir de 50 exécutions et suivant une loi t de Student.

Indicateurs	Modèles			
	VODUM	TAM	LDA	JTV
VIA moyenne	**0,639**	0,620	0,568	0,548
IC à 95 %	[0,609 ; 0,670]	[0,594 ; 0,647]	[0,551 ; 0,585]	[0,539 ; 0,558]
	VODUM-D	VODUM-W	VODUM-S	VODUM-O
VIA moyenne	0,598	0,517	0,564	0,518
IC à 95 %	[0,569 ; 0,626]	[0,513 ; 0,520]	[0,551 ; 0,577]	[0,514 ; 0,523]

VODUM-O et VODUM-W, respectivement – sont grandement bénéfiques à notre modèle pour identifier correctement le point de vue des documents. Ceci apporte donc confirmation aux hypothèses (H2) et (H3). Le modèle VODUM-S a obtenu une VIA supérieure à celles de VODUM-O et VODUM-W, mais néanmoins significativement plus faible que celle de VODUM. Ainsi, le positionnement des variables de point de vue au niveau des documents mène à une meilleure VIA que leur positionnement au niveau des phrases, soutenant (H4). Parmi les versions dégénérées, VODUM-D a globalement obtenu les meilleurs résultats, qui demeurent toutefois inférieurs à ceux de VODUM. Nous avons donc établi empiriquement la validité de la supposition faite dans [Qiu et Jiang, 2013; Qiu *et al.*, 2013b], postulant que l'utilisation de distributions de thèmes spécifiques aux points de vue (au lieu d'être spécifiques aux documents comme c'est le cas dans VODUM-D) améliore la VIA. Cela implique également la confirmation de l'hypothèse (H1).

En complément de l'analyse quantitative présentée dans cette section, nous étudions la qualité des thèmes, opinions et points de vue découverts par VODUM dans la Section 4.2.2.3.

4.2.2.3 Analyse qualitative des thèmes et des points de vue découverts par VODUM

L'analyse qualitative de notre modèle VODUM consiste à inspecter les mots thématiques et les mots d'opinions liés aux thèmes correspondant afin d'évaluer leur cohérence. Pour chaque thème z à évaluer, nous sélectionnons les 20 mots thématiques les plus probables dans la distribution indépendante des points de vue $\hat{\boldsymbol{\phi}}_{0z}$ et les 20 mots d'opinions les plus probables dans chaque distribution spécifique au point de vue $\hat{\boldsymbol{\phi}}_{1vz}$, $v = 1, \ldots, V$. Ces mots peuvent être considérés comme les plus représentatifs des thèmes et des points de vue identifiés par VODUM. Les distributions utilisées ici sont les estimateurs obtenus en sortie de l'échantillonneur de Gibbs marginalisé (Équations (4.5), (4.6), (4.7), (4.8)). En particulier, les résultats rapportés dans la suite de cette section proviennent de l'exécution qui a obtenu

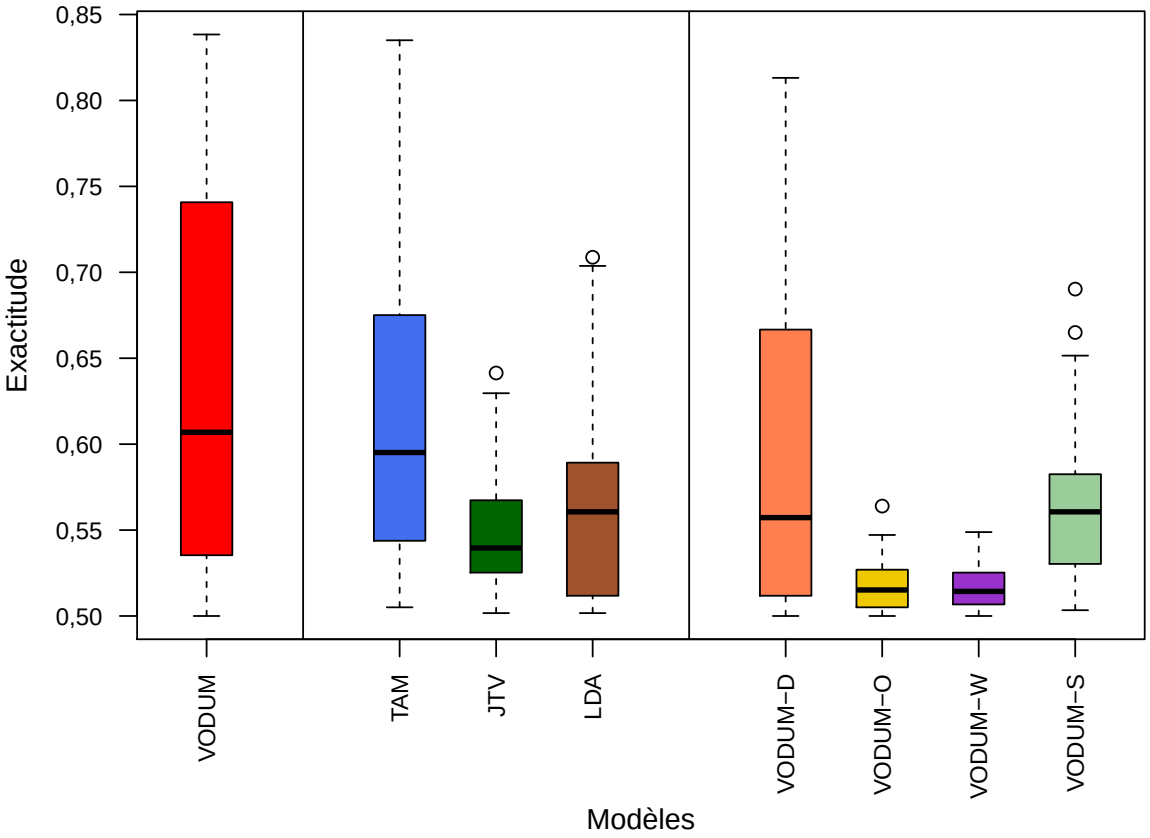


FIGURE 4.4 – Exactitude d'identification des points de vue (VIA) pour VODUM, TAM, JTV, LDA, VODUM-D, VODUM-O, VODUM-W et VODUM-S. Une valeur plus élevée indique une meilleure performance. Chaque boîte à moustache est calculée à partir de 50 exécutions.

la meilleure VIA dans la Section 4.2.2.2. Nous présentons dans les Tableaux 4.4 et 4.5 les listes de mots [11] thématiques et d'opinion les plus probables pour deux thèmes choisis, que l'on peut respectivement étiqueter comme « conflits au Moyen-Orient » et « justice et législation ». Nous indiquons également la probabilité de chacun de ces mots (deuxième, quatrième et sixième colonnes), fournie par les estimateurs $\hat{\phi}_0$ et $\hat{\phi}_1$.

Pour le thème des « conflits au Moyen-Orient » (Tableau 4.4), on peut observer que les mots thématiques (première colonne) sont globalement cohérents, avec des mots comme *syria*, *jihad*, *war* et *iraq*. Sans surprise, des mots comme *islam*, *terrorist* et *american* sont utilisés par le bord israélien (troisième colonne) pour évoquer les conflits du Moyen-Orient. Au contraire, le bord palestinien (cinquième colonne) reste très général vis-à-vis des conflits avec des mots comme *win*, *strong* et *commit*, sans associer les conflits au terrorisme et à l'islam.

11. Les mots sont en réalité des racines, obtenues par racinisation de Porter comme décrit dans la Section 4.2.1.2. Par conséquent, certaines transformations morphologiques ont été opérées – par exemple *dai* correspond à *day* et *days*, *suicid* correspond à *suicide* et *suicides*.

TABLEAU 4.4 – Listes des 20 mots thématiques et des 20 mots d'opinion (racinisés) les plus probables associés au thème manuellement étiqueté comme « conflits au Moyen-Orient ». Les mots d'opinions sont donnés pour chaque point de vue : israélien (I, $v = 1$) et palestinien (P, $v = 2$). À droite de chaque mot, nous indiquons sa probabilité, fournie par les estimateurs $\hat{\boldsymbol{\phi}}_0$ et $\hat{\boldsymbol{\phi}}_1$.

Mots thématiques	$\hat{\phi}_{0zw}$	Mots d'opinion (I)	$\hat{\phi}_{11zw}$	Mots d'opinion (P)	$\hat{\phi}_{12zw}$
israel	0,036	islam	0,051	need	0,032
palestinian	0,024	isra	0,031	win	0,024
syria	0,020	terrorist	0,025	think	0,024
jihad	0,013	recent	0,025	sai	0,024
war	0,012	militari	0,020	don	0,020
iraq	0,011	intern	0,017	strong	0,016
dai	0,011	like	0,014	new	0,016
suicid	0,011	heavi	0,011	sure	0,016
destruct	0,011	close	0,011	believ	0,016
iran	0,011	american	0,011	commit	0,012
fateh	0,011	certainli	0,011	person	0,012
iraqi	0,010	possibl	0,011	present	0,012
hama	0,010	radic	0,011	rule	0,012
man	0,010	capabl	0,011	achiev	0,012
saddam	0,009	nuclear	0,011	big	0,012
gaza	0,008	instead	0,011	versu	0,012
mass	0,008	polit	0,008	averag	0,012
bashar	0,008	stop	0,008	high	0,008
asad	0,008	pai	0,008	long	0,008
time	0,007	extrem	0,008	repres	0,008

Le thème de la « justice et protection » (Tableau 4.5) montre des résultats similaires. Les mots thématiques (première colonne) sont dans l'ensemble relativement cohérents avec des mots comme *law*, *court*, *justice*. On note toutefois que ce thème contient également des mots associés à la sécurité (*wall*, *secur*, *fenc*), ce qui indique le regroupement de plusieurs thèmes sous « justice et protection ». Le vocabulaire employé par le bord israélien (troisième colonne) est cette fois-ci général et peu informatif (*gener*, *advisori*, *accept*, *provid*, *polit*). Les mots utilisés par le bord palestinien (cinquième colonne) sont pour leur part plus spécifique avec la mention des droits fondamentaux et du secours humanitaire comme le montrent les mots *humanitarian*, *fundament*, *equit*.

En observant ces deux thèmes, nous notons la tendance suivante : différents thèmes semblent être privilégiés par différents points de vue. En effet, le thème sur les « conflits au Moyen-Orient » semblent plus fortement associés au point de vue israélien, alors que le thème de la « justice et protection » apparaît comme plus spécifique au point de vue palestinien. Dans ces deux cas, le bord peu concerné par un thème emploie un vocabulaire général qui indique une moins forte association entre les mots d'opinion relatifs à son point de vue

Tableau 4.5 – Listes des 20 mots thématiques et des 20 mots d'opinion (racinisés) les plus probables associés au thème manuellement étiqueté comme « justice et protection ». Les mots d'opinions sont donnés pour chaque point de vue : israélien (I, $v = 1$) et palestinien (P, $v = 2$). À droite de chaque mot, nous indiquons sa probabilité, fournie par les estimateurs $\boldsymbol{\phi}_0$ et $\boldsymbol{\phi}_1$.

Mots thématiques	$\hat{\phi}_{0zw}$	Mots d'opinion (I)	$\hat{\phi}_{11zw}$	Mots d'opinion (P)	$\hat{\phi}_{12zw}$
israel	0,077	intern	0,119	intern	0,138
law	0,058	isra	0,038	isra	0,059
court	0,052	high	0,033	legal	0,027
palestinian	0,037	legal	0,031	illeg	0,021
convent	0,033	gener	0,024	humanitarian	0,019
wall	0,023	advisori	0,021	west	0,015
secur	0,018	accept	0,014	militari	0,015
justic	0,017	provid	0,012	recent	0,015
territori	0,016	polit	0,012	fundament	0,013
water	0,016	current	0,012	specif	0,013
fenc	0,014	terribl	0,012	accord	0,013
geneva	0,014	appli	0,012	implement	0,013
hagu	0,014	right	0,010	relev	0,013
power	0,014	bring	0,010	appli	0,013
decis	0,013	deal	0,010	formal	0,013
rule	0,012	sai	0,010	equit	0,013
applic	0,012	illeg	0,010	high	0,010
opinion	0,010	begin	0,007	human	0,010
commun	0,010	think	0,007	occupi	0,010
occup	0,009	recent	0,007	natur	0,010

et les mots thématiques de ce thème. Par conséquent, cette observation soutient également l'hypothèse (H1).

4.3 Discussions

Nous avons présenté dans ce chapitre VODUM, un modèle thématique non supervisé permettant la découverte des points de vue, des thèmes et des mots thématiques et d'opinion associés dans une collection de documents pour lesquels aucune métadonnée n'est disponible. À travers les expérimentations réalisées, nous avons montré que notre modèle surpasse les approches de l'état de l'art dans le regroupement de points de vue. Nous avons également analysé l'impact des différentes spécificités de notre modèle. Le résultat de l'évaluation suggère que la séparation des mots d'opinion et des mots thématiques – basée sur les parties de discours – ainsi que l'utilisation d'attributions de thèmes au niveau des phrases, de points de vue au niveau des documents et de distributions de thèmes spécifiques aux points de vue améliore la capacité de notre modèle à regrouper correctement les points de vue. De plus,

l'étude qualitative que nous avons menée confirme la cohérence des mots thématiques et des mots d'opinion inférés par notre modèle.

Le travail présenté dans ce chapitre peut être étendu et amélioré sur plusieurs aspects. Comme nous l'avons évoqué dans la Section 4.2.2.2, notre modèle est sujet à une importante variance dans sa capacité à identifier correctement les points de vue. Cela pourrait être le signe d'un mélange lent de la chaîne de Markov associé à notre échantillonneur de Gibbs, probablement dû au fort couplage entre $\mathbf{v}$ et $\mathbf{z}$. Par conséquent, l'espace des variables $\mathbf{v}$ et $\mathbf{z}$ pourrait ne pas être exploré de manière satisfaisante en un nombre raisonnable d'itérations, amenant ainsi la chaîne de Markov dans un état ne reflétant pas la loi postérieure réelle des variables latentes. Ce type de difficultés a été mentionné dans un travail antérieur [Ahmed et Xing, 2010]. Pour parer ce problème, les auteurs ont ajouté une étape de Metropolis-Hasting à leur échantillonneur de Gibbs afin de faciliter l'exploration de l'espace des variables latentes et ainsi améliorer le mélange de la chaîne de Markov à construire. Nous envisageons de tester une approche similaire sur notre modèle VODUM.

D'autres améliorations pourraient être apportées à VODUM sur le plan technique. Par exemple, nous avons ici fixé les hyperparamètres α, β_0, β_1 et γ à des valeurs prédéfinies alors qu'il est possible d'échantillonner ces hyperparamètres en adoptant une posture bayésienne [Newman *et al.*, 2009]. Cela a pour effet de faciliter la généralisation d'un modèle à plusieurs collections, en éliminant la contrainte d'un choix d'hyperparamètres spécifiques à chaque collection. Bien que souvent négligés, les hyperparamètres peuvent avoir un impact non négligeable sur les performances d'un modèle, comme Wallach *et al.* [2009] l'ont montré. Par conséquent, nous respectons ce précepte dans notre contribution suivante, présentée dans le Chapitre 5. Similairement, il pourrait être intéressant d'explorer des modèles non paramétriques à la fois en terme de thèmes et de points de vue, afin de permettre au modèle de s'adapter de manière plus fidèle aux collections de textes traitées.

Du point de vue conceptuel, une des limites de VODUM est de ne pas permettre l'intégration de métadonnées potentiellement disponibles sur les documents de la collection, telles que des informations portant sur leur auteur. Ces métadonnées peuvent toutefois constituer des indicateurs critiques à l'identification du point de vue. Par exemple, si les documents étudiés sont postés sur des plateformes de réseaux sociaux telles que Twitter et Facebook, l'interaction des autres utilisateurs avec le contenu de l'auteur (telle que le *retweet*) peut s'avérer révélatrice de leur point de vue. Nous présentons ainsi dans le Chapitre 5 un modèle thématique permettant de découvrir le point de vue des utilisateurs de réseaux sociaux en exploitant conjointement leurs interactions et leur contenu généré.

Chapitre 5

Intégration des interactions sur les réseaux sociaux pour la découverte de points de vue

Sommaire

Ce chapitre introduit notre seconde contribution à la fouille de points de vue [Thonet *et al.*, 2017], dans laquelle nous nous intéressons en particulier aux points de vue exprimés sur les réseaux sociaux. Notre objectif est ici d'analyser dans quelle mesure l'utilisation des interactions entre utilisateurs, en outre de leur contenu textuel généré, est bénéfique à l'identification de leurs points de vue. Notre intuition concernant l'importance des interactions pour ce type de problèmes découle du principe d'homophilie – que nous avons précédemment défini dans la Section 2.1.2 – selon lequel les individus « similaires » (par exemple dans leurs opinions politiques) ont une plus forte propension à créer des liens. Nous présentons ainsi dans la Section 5.1 le modèle SNVDM (*Social Network Viewpoint Discovery Model*) qui exploite conjointement le contenu généré par les utilisateurs et leurs interactions pour modéliser sans supervision à la fois les points de vue et les thèmes qui leur sont associés. Afin de surmonter les cas où le réseau d'interactions sociales est peu dense (*sparse*) – lorsqu'un utilisateur n'interagit qu'avec un nombre limité d'utilisateurs tierces – nous proposons

dans la Section 5.2 une extension de SNVDM, nommée SNVDM-GPU, basée sur les urnes de Pólya généralisées. SNVDM-GPU présente notamment l'avantage d'intégrer les relations d'« accointances d'accointances » afin de prendre en compte les liens faibles existant entre les utilisateurs. La Section 5.3 décrit les expérimentations que nous avons menées sur deux collections de données issues de Twitter pour évaluer à la fois quantitativement et qualitativement nos modèles SNVDM et SNVDM-GPU, et les comparer aux différentes approches de l'état de l'art. La Section 5.4 clôture le chapitre et propose différentes extensions possibles aux modèles introduits.

5.1 SNVDM : un modèle thématique pour la découverte de points de vue dans les réseaux sociaux

Le modèle SNVDM, détaillé dans cette section, est une extension de LDA qui intègre à la fois thèmes et points de vue latents. La spécificité de ce modèle est de faire dépendre les points de vue non seulement du contenu textuel généré par les utilisateurs mais également des interactions sociales en ligne entre les utilisateurs. La Section 5.1.1 présente dans un premier temps des définitions destinées à expliciter le cadre d'application de notre modèle, ainsi qu'une description des notations adoptées dans le chapitre. Le modèle SNVDM est ensuite défini dans la Section 5.1.2. La Section 5.1.3 détaille la procédure d'inférence postérieure approchée de SNVDM. Enfin, nous évoquons dans la Section 5.1.4 les limites de SNVDM dans le cas où le réseau d'interactions sociales est peu dense – auxquelles nous chercherons ensuite à remédier dans la Section 5.2.

5.1.1 Préliminaires

Avant de décrire notre modèle, dont l'objectif est la modélisation jointe des *points de vue* et des *thèmes* dans les *réseaux sociaux*, nous allons dans un premier temps établir les définitions des termes clés. Les notions de points de vue et de thèmes ont déjà été abordées précédemment dans les Chapitres 2 et 3. Nous utilisons ici le terme « réseau social » pour dénoter le graphe dirigé comprenant les interactions sociales en ligne (les arêtes) entre les utilisateurs (les nœuds) sur une plateforme de médias sociaux (par exemple, Twitter). Par la suite, nous désignerons les interactions sociales en ligne simplement par *interactions sociales* ou *interactions*. Étant donnée la nature dirigée du réseau social, un utilisateur peut prendre part à deux types d'interactions distinctes : les *interactions entrantes* et les *interactions sortantes*. Un utilisateur u prend part à une interaction entrante avec un autre utilisateur u' si u' a initié l'interaction (par exemple, u' *retweete* ou répond au *tweet* de u). Nous appelons alors u et u' respectivement utilisateur cible (*recipient*) et utilisateur source (*sender*) de cette interaction. À l'inverse, un utilisateur u prend part à une interaction sortante avec un autre utilisateur u' si u a initié l'interaction (par exemple, u *retweete* ou répond au *tweet* de u'). Similairement, nous dénommons alors u et u' respectivement comme utilisateur source et utilisateur cible de l'interaction. Dans ces deux cas, nous dirons que u et u' constituent l'un

pour l'autre une *accointance en ligne*, ou simplement une *accointance*. En d'autres termes, l'ensemble des accointances d'un utilisateur u contient les utilisateurs sources des interactions entrantes sur u et les utilisateurs cibles des interactions sortantes de u.

Les notations des variables utilisées dans le reste de ce chapitre sont décrites dans le Tableau 5.1. Par ailleurs, nous utilisons comme précédemment des symboles en gras pour représenter les ensembles ou les vecteurs (par exemple, $\mathbf{v} = \left\{\{v_{ud}\}_{d=1}^{D_u}\right\}_{u=1}^{U}$). Nous rappelons que les compteurs, c'est-à-dire les variables exprimant le nombre d'attributions spécifiques, sont définies par n et des indices précisant les attributions (par exemple, n_{uz} désigne le nombre de mots de l'utilisateur u attribués au thème z). Un indice « $\bullet$ » utilisé à la place d'une variable indique que la marginalisation (sommation) du compteur pour toutes les valeurs prises par cette variable (par exemple, $n_{v\bullet} = \sum_{u=1}^{U} n_{vu}$). Un exposant $-(y)$ pour un ensemble de variables ou pour un compteur exclut y de l'ensemble ou du comptage (par exemple, $n_{uz}^{-(udn)}$ est le nombre de mots de l'utilisateur u assigné au thème z sans compter le n^{e} mot du document d de l'utilisateur u). Similairement, un exposant (y) pour un compteur inclut uniquement y lors du comptage (par exemple, $n_{uz}^{(ud)}$ est le nombre de mots de l'utilisateur u assigné au thème z dans le document d de l'utilisateur u).

5.1.2 Description du modèle

Avant de fournir une description formelle de notre modèle SNVDM (*Social Network Viewpoint Discovery Model*), nous en définissons les principales caractéristiques. Dans les modèles thématiques traditionnels tels que LDA [Blei *et al.*, 2003], les mots sont considérés comme étant tirés de distributions spécifiques aux thèmes. Cependant, dans SNVDM, nous avons pour objectif de modéliser à la fois les thèmes et les points de vue. Or, le point de vue d'un auteur peut influencer son choix de mots lors de la rédaction d'un texte – comme nous l'avions mentionné précédemment dans la Section 2.1.3 à travers des exemples que nous rappelons :

— Pendant la guerre des Six Jours, Israël a *occupé les territoires palestiniens* de la bande de Gaza.

— Pendant la guerre des Six Jours, Israël a *aménagé des implantations* dans la bande de Gaza.

Une approche naïve pour prendre en compte ce phénomène lexical consisterait alors à considérer que tous les mots sont tirés de distributions spécifiques à la fois aux thèmes et aux points de vue. Toutefois, tout mot n'est pas nécessairement relié à un thème et à un point de vue. Par exemple, un mot tel que *jours* peut être un « mot de fond » (*background word*) : il ne reflète ni point de vue, ni thème. Un mot peut également être thématique (par exemple, *guerre*) et ne dépendre que d'un thème, sans être spécifique à un quelconque point de vue. Pour intégrer cette observation à notre modèle, nous avons suivi l'idée proposée dans TAM (*Topic Aspect Model*) [Paul et Girju, 2010]. Les auteurs ont introduit des variables latentes de Bernoulli, nommées niveaux (*levels*) et routes (*routes*), afin de prendre en compte la possible dépendance des mots vis-à-vis des thèmes et des points de vue, respectivement. L'attribution d'un mot de la collection au niveau 0 indique que le mot ne reflète aucun thème et son attribution au niveau 1 indique au contraire qu'il est spécifique à un thème. Similairement,

Tableau 5.1 – Notations adoptées pour décrire notre modèle SNVDM.

Symbole	Définition
U, D_u, O_u, N_{ud}, I_{ud}	Nombre d'utilisateurs, nombre de documents de l'utilisateur u, nombre d'interactions sortantes pour l'utilisateur u, nombre de mots dans le document d de l'utilisateur u, nombre d'interactions entrantes pour le document d de l'utilisateur u, respectivement.
T, V, W	Nombre de thèmes, nombre de points de vue, taille du vocabulaire, respectivement.
w_{udn}, z_{udn}, ℓ_{udn}, x_{udn}	Le n^{e} mot du document d de l'utilisateur u et le thème, le niveau et la route qui lui sont attribués, respectivement.
s_{udi}	L'utilisateur source de l'interaction entrante i sur le document d de l'utilisateur u.
v_{ud}	Le point de vue attribué au document d de l'utilisateur u.
r_{uo}, v'_{uo}	L'utilisateur cible et le point de vue attribué pour l'interaction sortante o de l'utilisateur u, respectivement.
$\boldsymbol{\phi}_{00}$, $\boldsymbol{\phi}_{01}$, $\boldsymbol{\phi}_{10}$, $\boldsymbol{\phi}_{11}$, β	Matrice de taille $1 \times W$ contenant la distribution de mots de fond, matrice de taille $V \times W$ contenant les distributions de mots de point de vue, matrice de taille $T \times W$ contenant les distributions de mots thématiques, matrice de taille $V \times T \times W$ contenant les distributions de mots de point de vue thématiques, et leur paramètre de concentration, respectivement.
$\boldsymbol{\psi}_0$, $\boldsymbol{\psi}_1$, γ_0, γ_1	Matrice de taille 1×2 contenant la distribution de routes générale, matrice de taille $T \times 2$ contenant les distributions de routes spécifiques aux thèmes, et leurs paramètres de forme, respectivement.
$\boldsymbol{\xi}$, μ	Matrice de taille $V \times U$ contenant les distributions d'utilisateurs en interaction spécifiques aux points de vue, et leur paramètre de concentration, respectivement.
$\boldsymbol{\sigma}$, δ_0, δ_1	Matrice de taille $V \times U$ contenant les distributions de niveaux spécifiques aux utilisateurs, et leurs paramètres de forme, respectivement.
$\boldsymbol{\theta}$, α	Matrice de taille $U \times T$ contenant les distributions de thèmes spécifiques aux utilisateurs et leur paramètre de concentration, respectivement.
$\boldsymbol{\pi}$, η	Matrice de taille $U \times V$ contenant les distributions de points de vue spécifiques aux utilisateurs et leur paramètre de concentration, respectivement.
n_{00w}, n_{01vw}, n_{10zw}, n_{11vzw}	Nombre d'occurrences de w en tant que mot de fond, nombre d'occurrences de w en tant que mot de point de vue attribué à v, nombre d'occurrences de w en tant que mot thématique attribué à z, nombre d'occurrences de w en tant que mot de point de vue thématique attribué à v et z, respectivement.
n_{0x}, n_{1zx}	Nombre de mots assignés au niveau 0 et à la route x, nombre de mots assignés au niveau 1, au thème z et à la route x, respectivement.
n_{vu}	Nombre d'interactions avec l'accointance u attribuées au point de vue v.
$n_{u\ell}$	Nombre de mots de l'utilisateur u attribués au niveau ℓ.
n_{uz}	Nombre de mots de l'utilisateur u attribués au thème z.
n_{uv}	Nombre de documents et d'interactions sortantes de l'utilisateur u attribués au point de vue v.

une route 0 signifie l'indépendance du mot concerné vis-à-vis des différents points de vue et une route 1 exprime sa dépendance vis-à-vis d'un point de vue. On obtient ainsi quatre types de mots :

— les mots de fond – indépendants des thèmes et des points de vue (niveau = 0, route = 0) ;

— les mots de point de vue – dépendants uniquement des points de vue (niveau = 0, route = 1) ;

— les mots thématiques – dépendants uniquement des thèmes (niveau = 1, route = 0) ;

— les mots de point de vue thématiques – dépendants à la fois des thèmes et des points de vue (niveau = 1, route = 1).

Faire la distinction entre ces types de mots peut s'avérer très utile pour étudier l'utilisation du lexique spécifique aux points de vue ou au contraire indépendant de ceux-ci (lexique « neutre ») pour un thème donné.

Bien que le lexique aide partiellement à exprimer le point de vue dans un document, il peut ne pas suffire lorsque les points de vue sont exprimés de manière subtile. Ce problème est d'autant plus épineux pour les textes issus de médias sociaux, tels que les *tweets*, qui sont souvent courts, bruités et contiennent de nombreuses abréviations. Cependant, de nombreuses plateformes de médias sociaux permettent à leurs utilisateurs d'interagir. Le réseau social sous-jacent fournit alors de précieux indices sur les points de vue des utilisateurs en accord avec le principe d'homophilie, défini dans la Section 2.1.3. Pour rappel, l'homophilie postule que les personnes « semblables » ont tendance à créer des liens entre elles. Dans notre cas, cela signifie que les utilisateurs avec des points de vue similaires vont avoir une plus grande propension à interagir les uns avec les autres.

L'idée d'exploiter le phénomène d'homophilie a déjà été abordée dans des modèles thématiques antérieurs, construits pour la détection de communautés – dont nous avons donné un aperçu dans la Section 2.3.2.2. Un exemple de tels modèles exploitant les interactions entre utilisateurs sur les réseaux sociaux est SN-LDA (*Social Network Latent Dirichlet Allocation*) [Sachan *et al.*, 2014]. Cependant, SN-LDA se base uniquement sur les interactions et les attributions latentes de thèmes pour identifier les communautés des utilisateurs – sans supposer directement que différentes communautés adoptent des lexiques différents pour décrire les mêmes thèmes. Cela s'explique par le fait que les auteurs de SN-LDA s'intéressent à la détection de communautés dites « thématiques » – c'est-à-dire, qui parlent de thèmes différents (par exemple, la *programmation informatique* ou le *football*). Dans notre cas cependant, les différentes communautés – chacune étant associée à un point de vue – s'expriment sur un ensemble de thèmes communs (par exemple, la *protection sociale* et l'*environnement*, dans le cas des points de vue politiques).

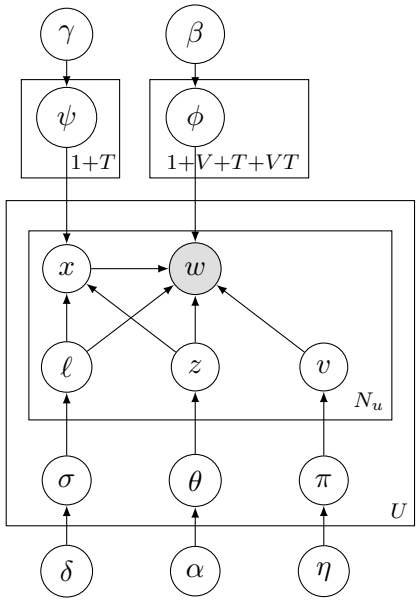
De plus dans SN-LDA, la communauté (ou le rattachement à un point de vue) d'un utilisateur donné est seulement influencée par ses interactions « sortantes » (*outgoing interactions*), c'est-à-dire les interactions qu'il a initiées – telles que l'action de *retweeter* un autre utilisateur. Les interactions « entrantes » (*incoming interactions*), actions que l'utilisateur a subies – telles que le fait d'être *retweeté* par un autre utilisateur, sont elles ignorées. Ces

dernières interactions peuvent néanmoins être riches en information pour un utilisateur qui est connecté principalement ou uniquement par des interactions entrantes (par exemple dans le cas d'une personnalité qui est souvent *retweeté* grâce à sa popularité, mais qui ne *retweete* que rarement). Suivant la modélisation faite par SN-LDA, la communauté d'un tel utilisateur sera difficile à identifier en raison de son faible nombre d'interactions sortantes.

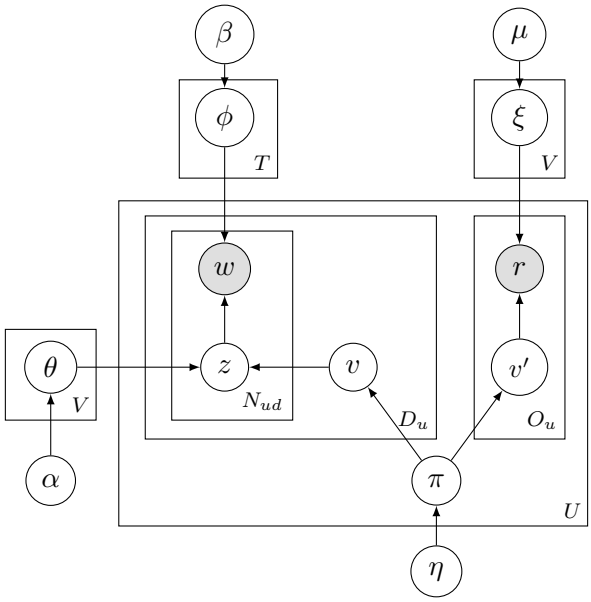
À l'inverse, dans notre modèle SNVDM, nous proposons d'utiliser à la fois les *interactions entrantes et sortantes*. Plus précisément, dans SNVDM, les interactions entrantes sont exploitées au niveau du document pour dénoter l'influence rétrospective des utilisateurs en interaction avec le document (par exemple, les utilisateurs *retweetant* un *tweet* donné) sur le point de vue exprimé dans le document. Notons que nous adoptons ici une vision statique de la collection de documents, nous supposerons par conséquent dans l'histoire générative de SNVDM que les interactions entrantes sont effectuées immédiatement après rédaction du document. En d'autres termes, nous ne tenons pas compte de la temporalité des interactions entrantes.

Nous apportons maintenant une définition plus formelle de notre modèle SNVDM en décrivant le modèle graphique associé (Figure 5.1c) comparé à celui des modèles TAM (Figure 5.1a) et SN-LDA (Figure 5.1b) dont nous nous sommes inspirés. L'histoire générative de SNVDM est la suivante :

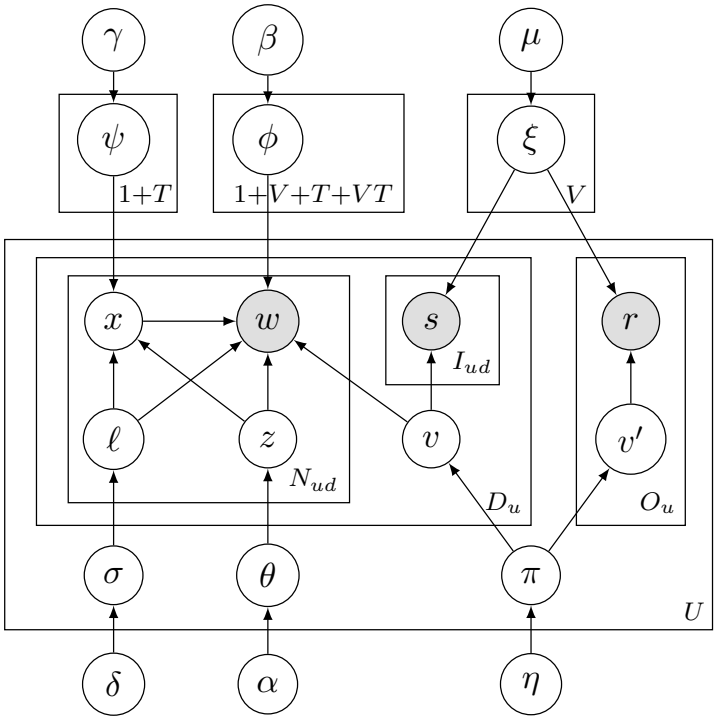
1. Tirer les distributions de routes générales et spécifiques aux thèmes $\boldsymbol{\psi}_0$ et $\boldsymbol{\psi}_{1z} \sim \text{Beta}(\gamma_0, \gamma_1)$ pour $z = 1, ..., T$, respectivement ;
2. Tirer les distributions de mots de fond, de mots de point de vue, de mots thématiques et de mots de point de vue thématiques $\boldsymbol{\phi}_{00}$, $\boldsymbol{\phi}_{01v}$, $\boldsymbol{\phi}_{10z}$ et $\boldsymbol{\phi}_{11vz} \sim \text{Dirichlet}_W(\beta)$ pour $v = 1, ..., V$ et $z = 1, ..., T$, respectivement ;
3. Tirer les distributions d'utilisateurs en interaction spécifiques aux points de vue $\boldsymbol{\xi}_v \sim \text{Dirichlet}_U(\mu\frac{1}{U})$ pour $v = 1, ..., V$;
4. Pour chaque utilisateur $u = 1, ..., U$:
   (a) Tirer une distribution de points de vue $\boldsymbol{\pi}_u \sim \text{Dirichlet}_V(\eta\frac{1}{V})$;
   (b) Tirer une distribution de thèmes $\boldsymbol{\theta}_u \sim \text{Dirichlet}_T(\alpha\frac{1}{T})$;
   (c) Tirer une distribution de niveaux $\boldsymbol{\sigma}_u \sim \text{Beta}(\delta_0, \delta_1)$;
5. Pour chaque document $d = 1, ..., D_u$ de $u = 1, ..., U$:
   (a) Tirer le point de vue du document $v_{ud} \sim \text{Discrète}(\boldsymbol{\pi}_u)$;
   (b) Pour chaque emplacement de mots $n = 1, ..., N_{ud}$:
      i. Tirer un thème $z_{udn} \sim \text{Discrète}(\boldsymbol{\theta}_u)$;
      ii. Tirer un niveau $\ell_{udn} \sim \text{Bernoulli}(\boldsymbol{\sigma}_u)$;
      iii. Si $\ell_{udn} = 0$,
         Tirer une route générale $x_{udn} \sim \text{Bernoulli}(\boldsymbol{\psi}_0)$;
      Sinon si $\ell_{udn} = 1$,
         Tirer une route spécifique au thème $x_{udn} \sim \text{Bernoulli}(\boldsymbol{\psi}_{1z_{udn}})$;



(a) TAM.



(b) SN-LDA.



(c) SNVDM.

FIGURE 5.1 – Représentation sous forme de modèle graphique de TAM (a), SN-LDA (b) et notre modèle SNVDM (c). Le nom des variables de TAM et SN-LDA a été légèrement modifié par rapport aux notations originales utilisées respectivement dans [Paul et Girju, 2010] et [Sachan *et al.*, 2014] afin de rendre la comparaison avec SNVDM plus aisée.

iv. Si $\ell_{udn} = 0$ et $x_{udn} = 0$,

Tirer un mot de fond $w_{udn} \sim$ Discrète($\boldsymbol{\phi}_{00}$) ;

Sinon si $\ell_{udn} = 0$ et $x_{udn} = 1$,

Tirer un mot de point de vue $w_{udn} \sim$ Discrète($\boldsymbol{\phi}_{01v_{ud}}$) ;

Sinon si $\ell_{udn} = 1$ et $x_{udn} = 0$,

Tirer un mot thématique $w_{udn} \sim$ Discrète($\boldsymbol{\phi}_{10z_{udn}}$) ;

Sinon si $\ell_{udn} = 1$ et $x_{udn} = 1$,

Tirer un mot de point de vue thématique $w_{udn} \sim$ Discrète($\boldsymbol{\phi}_{11v_{ud}z_{udn}}$) ;

(c) Pour chaque interaction entrante $i = 1, ..., I_{ud}$:
Tirer un utilisateur source $s_{udi} \sim$ Discrète($\boldsymbol{\xi}_{v_{ud}}$) ;

6. Pour chaque interaction sortante $o = 1, ..., O_u$ de $u = 1, ..., U$:

(a) Tirer le point de vue de l'interaction sortante $v'_{uo} \sim$ Discrète($\boldsymbol{\pi}_u$) ;

(b) Tirer un utilisateur cible $r_{uo} \sim$ Discrète($\boldsymbol{\xi}_{v'_{uo}}$).

L'expression complète de la probabilité jointe des variables latentes (attributions $\mathbf{x}$, $\boldsymbol{\ell}$, $\mathbf{z}$, $\mathbf{v}$, $\mathbf{v}'$, et distributions $\boldsymbol{\phi}$, $\boldsymbol{\psi}$, $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$, $\boldsymbol{\pi}$, $\boldsymbol{\xi}$) et observées ($\mathbf{w}$, $\mathbf{s}$, $\mathbf{r}$) du modèle est la suivante :

$$
\begin{aligned}
& p(\mathbf{w}, \mathbf{x}, \boldsymbol{\ell}, \mathbf{z}, \mathbf{s}, \mathbf{v}, \mathbf{r}, \mathbf{v}', \boldsymbol{\phi}, \boldsymbol{\psi}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\xi} \,;\, \beta, \gamma_0, \gamma_1, \delta_0, \delta_1, \alpha, \eta, \mu) \qquad (5.1) \\
= & \left\{ \prod_{u=1}^{U} \prod_{d=1}^{D_u} \prod_{n=1}^{N_{ud}} \theta_{uz_{udn}} \sigma_{u\ell_{udn}} \left[\psi_{0x_{udn}}\right]^{\mathbb{I}(\ell_{udn}=0)} \left[\psi_{1z_{udn}x_{udn}}\right]^{\mathbb{I}(\ell_{udn}=1)} \left[\phi_{00w_{udn}}\right]^{\mathbb{I}(\ell_{udn}=0 \wedge x_{udn}=0)} \right. \\
& \left. \left[\phi_{01v_{ud}w_{udn}}\right]^{\mathbb{I}(\ell_{udn}=0 \wedge x_{udn}=1)} \left[\phi_{10z_{udn}w_{udn}}\right]^{\mathbb{I}(\ell_{udn}=1 \wedge x_{udn}=0)} \left[\phi_{11v_{ud}z_{udn}w_{udn}}\right]^{\mathbb{I}(\ell_{udn}=1 \wedge x_{udn}=1)} \right\} \\
& \left\{ \prod_{u=1}^{U} \prod_{d=1}^{D_u} \pi_{uv_{ud}} \prod_{i=1}^{I_{ud}} \xi_{v_{ud}s_{udi}} \right\} \left\{ \prod_{u=1}^{U} \prod_{o=1}^{O_u} \pi_{uv'_{uo}} \xi_{v'_{uo}r_{uo}} \right\} \mathrm{Dir}_W(\boldsymbol{\phi}_{00z} \,;\, \beta) \left\{ \prod_{v=1}^{V} \mathrm{Dir}_W(\boldsymbol{\phi}_{01v} \,;\, \beta) \right\} \\
& \left\{ \prod_{z=1}^{T} \mathrm{Dir}_W(\boldsymbol{\phi}_{10z} \,;\, \beta) \right\} \left\{ \prod_{v=1}^{V} \prod_{z=1}^{T} \mathrm{Dir}_W(\boldsymbol{\phi}_{11vz} \,;\, \beta) \right\} \mathrm{Beta}(\boldsymbol{\psi}_0 \,;\, \gamma_0, \gamma_1) \left\{ \prod_{z=1}^{T} \mathrm{Beta}(\boldsymbol{\psi}_{1z} \,;\, \gamma_0, \gamma_1) \right\} \\
& \left\{ \prod_{u=1}^{U} \mathrm{Beta}(\boldsymbol{\sigma}_u \,;\, \delta_0, \delta_1) \, \mathrm{Dir}_T(\boldsymbol{\theta}_u \,;\, \alpha) \, \mathrm{Dir}_V(\boldsymbol{\pi}_u \,;\, \eta) \right\} \left\{ \prod_{z=1}^{T} \mathrm{Dir}_W(\boldsymbol{\xi} \,;\, \mu) \right\}
\end{aligned}
$$

où $\wedge$ désigne un « et » logique entre deux expressions booléennes. Comme c'est le cas pour les autres modèles thématiques basés sur LDA, l'inférence postérieure de SNVDM ne peut être effectuée de manière exacte. Nous décrivons dans la Section 5.1.3 notre procédure d'inférence approchée basée sur un échantillonneur de Gibbs marginalisé [Liu, 1994] – précédemment décrite dans Section 3.3.1.

5.1.3 Inférence postérieure

Dans le cas de SNVDM, les variables latentes d'intérêt pour un échantillonneur de Gibbs marginalisé sont les thèmes $\mathbf{z}$, les niveaux $\boldsymbol{\ell}$, les routes $\mathbf{x}$, les points de vue des documents $\mathbf{v}$

et les points de vue des interactions sortantes $\mathbf{v}'$. Les variables observées correspondent aux mots $\mathbf{w}$, aux utilisateurs sources des interactions entrantes $\mathbf{s}$ et aux utilisateurs cibles des interactions sortantes $\mathbf{r}$. L'échantillonneur de Gibbs marginalisé de SNVDM consiste alors à effectuer des tirages successifs des variables latentes précédentes à partir de leur probabilité conditionnelle postérieure – que nous indiquons dans le reste de cette section. Notons que pour éviter d'alourdir les formules exprimant une probabilité conditionnelle, nous dénotons par « reste » les variables aléatoires et les hyperparamètres omis.

Tirage des thèmes, niveaux et routes. De manière similaire à TAM [Paul et Girju, 2010], les attributions de thème z_{udn}, de niveau ℓ_{udn} et de route x_{udn} pour le n^{e} mot du document d de l'utilisateur u peuvent être conjointement tirés à partir de la probabilité postérieure suivante :

$$
\begin{aligned}
&p(z_{udn}=z,\ell_{udn}=\ell,x_{udn}=x|v_{ud}=v,w_{udn}=w,\text{reste}) \qquad (5.2)\\
&\propto \frac{p(\mathbf{z})}{p(\mathbf{z}^{-(udn)})}\cdot\frac{p(\boldsymbol{\ell})}{p(\boldsymbol{\ell}^{-(udn)})}\cdot\frac{p(\mathbf{x}|\mathbf{z},\boldsymbol{\ell})}{p(\mathbf{x}^{-(udn)}|\mathbf{z}^{-(udn)},\boldsymbol{\ell}^{-(udn)})}\cdot\frac{p(\mathbf{w}|\mathbf{v},\mathbf{z},\boldsymbol{\ell},\mathbf{x})}{p(\mathbf{w}^{-(udn)}|\mathbf{v},\mathbf{z}^{-(udn)},\boldsymbol{\ell}^{-(udn)},\mathbf{x}^{-(udn)})}\\
&=\frac{\Gamma(n_{u\bullet}^{-(udn)}+\alpha)}{\Gamma(n_{u\bullet}+\alpha)}\cdot\frac{\Gamma(n_{uz}+\alpha\frac{1}{T})}{\Gamma(n_{uz}^{-(udn)}+\alpha\frac{1}{T})}\cdot\frac{\Gamma(n_{u\bullet}^{-(udn)}+\delta_0+\delta_1)}{\Gamma(n_{u\bullet}+\delta_0+\delta_1)}\cdot\frac{\Gamma(n_{u\ell}+\delta_\ell)}{\Gamma(n_{u\ell}^{-(udn)}+\delta_\ell)}\\
&\quad\cdot\begin{cases}
\frac{\Gamma(n_{0\bullet}^{-(udn)}+\gamma_0+\gamma_1)}{\Gamma(n_{0\bullet}+\gamma_0+\gamma_1)}\cdot\frac{\Gamma(n_{0x}+\gamma_x)}{\Gamma(n_{0x}^{-(udn)}+\gamma_x)}\cdot\frac{\Gamma(n_{00\bullet}^{-(udn)}+\beta W)}{\Gamma(n_{00\bullet}+\beta W)}\cdot\frac{\Gamma(n_{00w}+\beta)}{\Gamma(n_{00w}^{-(udn)}+\beta)} & \text{si } \ell=0,\ x=0,\\
\frac{\Gamma(n_{0\bullet}^{-(udn)}+\gamma_0+\gamma_1)}{\Gamma(n_{0\bullet}+\gamma_0+\gamma_1)}\cdot\frac{\Gamma(n_{0x}+\gamma_x)}{\Gamma(n_{0x}^{-(udn)}+\gamma_x)}\cdot\frac{\Gamma(n_{01v\bullet}^{-(udn)}+\beta W)}{\Gamma(n_{01v\bullet}+\beta W)}\cdot\frac{\Gamma(n_{01vw}+\beta)}{\Gamma(n_{01vw}^{-(udn)}+\beta)} & \text{si } \ell=0,\ x=1,\\
\frac{\Gamma(n_{1z\bullet}^{-(udn)}+\gamma_0+\gamma_1)}{\Gamma(n_{1z\bullet}+\gamma_0+\gamma_1)}\cdot\frac{\Gamma(n_{1zx}+\gamma_x)}{\Gamma(n_{1zx}^{-(udn)}+\gamma_x)}\cdot\frac{\Gamma(n_{10z\bullet}^{-(udn)}+\beta W)}{\Gamma(n_{10z\bullet}+\beta W)}\cdot\frac{\Gamma(n_{10zw}+\beta)}{\Gamma(n_{10zw}^{-(udn)}+\beta)} & \text{si } \ell=1,\ x=0,\\
\frac{\Gamma(n_{1z\bullet}^{-(udn)}+\gamma_0+\gamma_1)}{\Gamma(n_{1z\bullet}+\gamma_0+\gamma_1)}\cdot\frac{\Gamma(n_{1zx}+\gamma_x)}{\Gamma(n_{1zx}^{-(udn)}+\gamma_x)}\cdot\frac{\Gamma(n_{11vz\bullet}^{-(udn)}+\beta W)}{\Gamma(n_{11vz\bullet}+\beta W)}\cdot\frac{\Gamma(n_{11vzw}+\beta)}{\Gamma(n_{11vzw}^{-(udn)}+\beta)} & \text{si } \ell=1,\ x=1\,;
\end{cases}\\
&=\frac{n_{uz}^{-(udn)}+\alpha\frac{1}{T}}{n_{u\bullet}^{-(udn)}+\alpha}\cdot\frac{n_{u\ell}^{-(udn)}+\delta_\ell}{n_{u\bullet}^{-(udn)}+\delta_0+\delta_1}\cdot\begin{cases}
\frac{n_{0x}^{-(udn)}+\gamma_x}{n_{0\bullet}^{-(udn)}+\gamma_0+\gamma_1}\cdot\frac{n_{00w}^{-(udn)}+\beta}{n_{00\bullet}^{-(udn)}+\beta W} & \text{si } \ell=0,\ x=0,\\
\frac{n_{0x}^{-(udn)}+\gamma_x}{n_{0\bullet}^{-(udn)}+\gamma_0+\gamma_1}\cdot\frac{n_{01vw}^{-(udn)}+\beta}{n_{01v\bullet}^{-(udn)}+\beta W} & \text{si } \ell=0,\ x=1,\\
\frac{n_{1zx}^{-(udn)}+\gamma_x}{n_{1z\bullet}^{-(udn)}+\gamma_0+\gamma_1}\cdot\frac{n_{10zw}^{-(udn)}+\beta}{n_{10z\bullet}^{-(udn)}+\beta W} & \text{si } \ell=1,\ x=0,\\
\frac{n_{1zx}^{-(udn)}+\gamma_x}{n_{1z\bullet}^{-(udn)}+\gamma_0+\gamma_1}\cdot\frac{n_{11vzw}^{-(udn)}+\beta}{n_{11vz\bullet}^{-(udn)}+\beta W} & \text{si } \ell=1,\ x=1.
\end{cases}
\end{aligned}
$$

Cette dérivation suit un schéma similaire à celle de l'échantillonneur de Gibbs marginalisé de LDA, fournie dans l'Équation (3.12) de la Section 3.3.1.3. Explicitons les termes obtenus dans la formule finale :

— Le premier terme reflète la fréquence de l'attribution du thème z aux mots des documents écrits par l'utilisateur u.

— Le deuxième terme indique la propension des mots des documents de l'utilisateur u à se voir attribuer le niveau ℓ. Si $\ell=0$ pour un mot donné, cela signifie que le mot et la

route associés ne dépendent pas du thème ; si $\ell = 1$, le mot et la route dépendent du thème.

— Le troisième terme exprime dans quelle mesure la route x est fréquente dans la collection toute entière. Si $x = 0$ est attribué à un mot, ce mot est indépendant du point de vue ; si $x = 1$, le mot reflète un point de vue.

— Le quatrième terme mesure la force de l'association entre le mot observé et les attributions de thèmes et/ou de points de vue.

Tirage des points de vue des documents. La probabilité conditionnelle postérieure de l'attribution de point de vue v_{ud} pour le document d écrit par l'utilisateur u s'obtient ainsi :

$$
\begin{aligned}
& p(v_{ud} = v | \mathbf{v}^{-(ud)}, \mathbf{s}, \mathbf{w}, \text{reste}) \qquad (5.3)\\
\propto\ & \frac{p(\mathbf{v})}{p(\mathbf{v}^{-(ud)})} \cdot \frac{p(\mathbf{s}|\mathbf{v})}{p(\mathbf{s}^{-(ud)}|\mathbf{v}^{-(ud)})} \cdot \frac{p(\mathbf{w}|\mathbf{v}, \mathbf{z}, \boldsymbol{\ell}, \mathbf{x})}{p(\mathbf{w}^{-(ud)}|\mathbf{v}^{-(ud)}, \mathbf{z}^{-(ud)}, \boldsymbol{\ell}^{-(ud)}, \mathbf{x}^{-(ud)})} \\
\propto\ & \frac{\Gamma(n_{u\bullet}^{-(ud)} + \eta)}{\Gamma(n_{u\bullet} + \eta)} \cdot \frac{\Gamma(n_{uv} + \eta\frac{1}{V})}{\Gamma(n_{uv}^{-(ud)} + \eta\frac{1}{V})} \cdot \frac{\Gamma(n_{v\bullet}^{-(ud)} + \mu)}{\Gamma(n_{v\bullet} + \mu)} \cdot \prod_{u'=1}^{U} \frac{\Gamma(n_{vu'} + \mu\frac{1}{U})}{\Gamma(n_{vu'}^{-(ud)} + \mu\frac{1}{U})} \\
& \cdot \frac{\Gamma(n_{01v\bullet}^{-(ud)} + \beta W)}{\Gamma(n_{01v\bullet} + \beta W)} \cdot \prod_{w=1}^{W} \frac{\Gamma(n_{01vw} + \beta)}{\Gamma(n_{01vw}^{-(ud)} + \beta)} \cdot \prod_{z=1}^{T} \left\{ \frac{\Gamma(n_{11vz\bullet}^{-(ud)} + \beta W)}{\Gamma(n_{11vz\bullet} + \beta W)} \cdot \prod_{w=1}^{W} \frac{\Gamma(n_{11vzw} + \beta)}{\Gamma(n_{11vzw}^{-(ud)} + \beta)} \right\} \\
=\ & \frac{n_{uv}^{-(ud)} + \eta\frac{1}{V}}{n_{u\bullet}^{-(ud)} + \eta} \cdot \frac{\prod_{u'=1}^{U} \prod_{a=0}^{n_{vu'}^{(ud)}-1} (n_{vu'}^{-(ud)} + a + \mu\frac{1}{U})}{\prod_{b=0}^{n_{v\bullet}^{(ud)}-1} (n_{v\bullet}^{-(ud)} + b + \mu)} \\
& \cdot \frac{\prod_{w=1}^{W} \prod_{a=0}^{n_{01vw}^{(ud)}-1} (n_{01vw}^{-(ud)} + a + \beta)}{\prod_{b=0}^{n_{01v\bullet}^{(ud)}-1} (n_{01v\bullet}^{-(ud)} + b + \beta W)} \cdot \prod_{z=1}^{T} \frac{\prod_{w=1}^{W} \prod_{a=0}^{n_{11vzw}^{(ud)}-1} (n_{11vzw}^{-(ud)} + a + \beta)}{\prod_{b=0}^{n_{11vz\bullet}^{(ud)}-1} (n_{11vz\bullet}^{-(ud)} + b + \beta W)}.
\end{aligned}
$$

où nous avons de nouveau utilisé l'identité $\Gamma(x + n) = \Gamma(x) \prod_{i=0}^{n-1}(x + i)$ pour $x \in \mathbb{R}$ et $n \in \mathbb{N}$, introduite dans la Section 4.1.2. La probabilité postérieure de v_{ud} présente les termes suivants :

— Le premier terme dénote la fréquence de l'attribution du point de vue v aux documents et aux interactions sortantes de l'utilisateur u (en tant qu'auteur et utilisateur source).

— Le deuxième terme indique la propension des utilisateurs sources d'interactions entrantes sur le document d de l'utilisateur u à se voir attribuer le point de vue v dans leurs autres interactions entrantes ou sortantes.

— Les troisième et quatrième termes expriment dans quelle mesure respectivement les mots de point de vue et les mots de point de vue thématiques du document d de l'utilisateur u apparaissent également dans des documents assignés à v dans le reste de la collection.

Tirage des points de vue d'interactions sortantes. Pour chaque interaction sortante o de l'utilisateur u, le point de vue v'_{uo} est tiré selon la probabilité conditionnelle postérieure

suivante :

$$p(v'_{uo} = v | r_{uo} = u', \mathbf{r}^{-(uo)}, \mathbf{v'}^{-(uo)}, \text{reste}) \tag{5.4}$$

$$\propto \frac{p(\mathbf{v'})}{p(\mathbf{v'}^{-(uo)})} \cdot \frac{p(\mathbf{r}|\mathbf{v'})}{p(\mathbf{r}^{-(uo)}|\mathbf{v'}^{-(uo)})}$$

$$= \frac{\Gamma(n_{u\bullet}^{-(uo)} + \eta)}{\Gamma(n_{u\bullet} + \eta)} \cdot \frac{\Gamma(n_{uv} + \eta\frac{1}{V})}{\Gamma(n_{uv}^{-(uo)} + \eta\frac{1}{V})} \cdot \frac{\Gamma(n_{v\bullet}^{-(uo)} + \mu)}{\Gamma(n_{v\bullet} + \mu)} \cdot \frac{\Gamma(n_{vu'} + \mu\frac{1}{U})}{\Gamma(n_{vu'}^{-(uo)} + \mu\frac{1}{U})}$$

$$= \frac{n_{uv}^{-(uo)} + \eta\frac{1}{V}}{n_{u\bullet}^{-(uo)} + \eta} \cdot \frac{n_{vu'}^{-(uo)} + \mu\frac{1}{U}}{n_{v\bullet}^{-(uo)} + \mu}. \tag{5.5}$$

où l'on retrouve les termes suivants :

— Le premier terme correspond de nouveau à la fréquence de l'attribution du point de vue v aux documents et aux interactions sortantes de l'utilisateur u (en tant qu'auteur et utilisateur source).

— Le second terme dénote la propension de l'utilisateur cible u' (associé à l'interaction sortante o de l'utilisateur u) à se voir attribuer le point de vue v dans ses autres interactions entrantes ou sortantes.

Tirage des hyperparamètres. Il a été montré par Wallach *et al.* [2009] que le tirage des hyperparamètres suivant leur probabilité postérieure – plutôt que l'utilisation de valeurs fixées – impacte positivement les performances des modèles thématiques. Par conséquent, nous adoptons ici une posture bayésienne sur la modélisation des hyperparamètres en imposant un *hyperprior* Gamma$(1, 1)$ à α, γ_0, γ_1, δ_0, δ_1, η et μ. L'échantillonnage des hyperparamètres est alors effectué en s'appuyant sur la technique dite de « variables auxiliaires » (*auxiliary variable sampling*) proposée par Escobar et West [1995] et appliquée aux modèles thématiques dans [Newman *et al.*, 2009; Teh *et al.*, 2006]. Nous fixons seulement l'hyperparamètre $\beta = 0.01$ car son échantillonnage est plus coûteux – étant donné qu'il dépend de la taille du vocabulaire, qui est généralement large.

5.1.4 Limites de SNVDM

Bien que SNVDM utilise les interactions entrantes et sortantes directes entre un utilisateur et ses accointances en ligne, notre modèle n'exploite pas les liens d'ordre supérieur entre utilisateurs tels que les *accointances d'accointances* – qui peuvent toutefois constituer des indices précieux pour identifier le point de vue des utilisateurs. Par exemple, considérons un utilisateur u qui interagit exclusivement avec un second utilisateur u', lui-même interagissant avec un grand nombre d'utilisateurs. Utiliser seulement les rares interactions de u (en plus de son contenu généré) ne fournira que très peu d'informations au sujet de cet utilisateur. Néanmoins, savoir qu'il existe un lien faible (du deuxième ordre) entre u et les accointances de u' peut aider à décrire u plus précisément. Nous étendons SNVDM dans la Section 5.2 pour mettre en œuvre cette idée.

5.2 SNVDM-GPU : extension de SNVDM basée sur les urnes de Pólya généralisées

Nous proposons d'étendre SNVDM afin de prendre en compte les liens distants entre les utilisateurs et leurs « accointances d'accointances ». Nous introduisons dans un premier temps la processus associé aux urnes de Pólya simples (Section 5.2.1) et aux urnes de Pólya généralisées (Section 5.2.2), puis nous décrivons notre extension de SNVDM basée sur les urnes de Pólya généralisées (Section 5.2).

5.2.1 Urnes de Pólya simples

L'échantillonneur de Gibbs marginalisé pour les modèles thématiques exploitant la relation de conjugaison entre distributions de Dirichlet et distributions multinomiales (tels que LDA et SNVDM) peut être interprété selon le processus d'urne de Pólya (*Pólya urn process* ou *Pólya urn scheme*). Nous dénommerons ce dernier « processus d'urne de Pólya simple » (SPU – *simple Pólya urn*) afin de le distinguer de sa version généralisée que nous décrirons dans la Section 5.2.2. Le processus de SPU est illustré dans la Figure 5.2. Le principe est le suivant : on tire successivement et aléatoirement des boules colorées d'une urne. Si une boule de couleur c est tirée, alors cette boule est remise dans l'urne en ajoutant également une boule supplémentaire de la même couleur c [1]. Ce processus avec « sur-remise » (*over-replacement*) – c'est-à-dire dans lequel on remet dans l'urne plus de boules que l'on en a tirées – met en application une propriété connue comme « *the rich get richer* » : plus les boules d'une couleur c sont tirées, plus la chance de tirer de nouveau des boules de couleur c à l'avenir est augmentée.

Nous appliquons maintenant la métaphore de SPU à SNVDM en supposant que chaque interaction sortante est une boule, que l'utilisateur cible de l'interaction est sa couleur et que chaque urne est associée à un point de vue. Initialement, chaque urne $v = 1, ..., V$ contient $\mu\frac{1}{U}$ boules [2] de chaque couleur. Nous supposons maintenant que l'interaction o de l'utilisateur u est la dernière boule à tirer de l'urne v – c'est-à-dire que nous avons déjà tiré, observé et (sur-)remis dans leurs urnes respectives toutes les autres interactions/boules de l'ensemble des utilisateurs. La probabilité d'obtenir la couleur u' pour la boule (uo) est alors donnée par la formule suivante :

$$p(r_{uo} = u' | r^{-(uo)}, v'_{uo} = v, \mathbf{v'}^{-(uo)}, \text{reste}) = \frac{n_{vu'}^{-(uo)} + \mu\frac{1}{U}}{n_{v\bullet}^{-(uo)} + \mu} \tag{5.6}$$

où $n_{vu'}^{-(uo)}$ peut être interprété comme le nombre de boules de couleur u' ajoutées précédemment à l'urne v et $n_{v\bullet}^{-(uo)}$ comme le nombre total de boules ajoutées précédemment à v, ces deux compteurs excluant la boule (uo) et les μ boules initiales. Cette formule est très intuitive :

1. Pour rendre l'expérience plus cohérente, nous supposons disposer d'une source infinie de boules colorées, représentée par le générateur sur la droite de la Figure 5.2.
2. Bien qu'en pratique ce nombre ne soit généralement pas entier, cela n'altère pas notre propos.

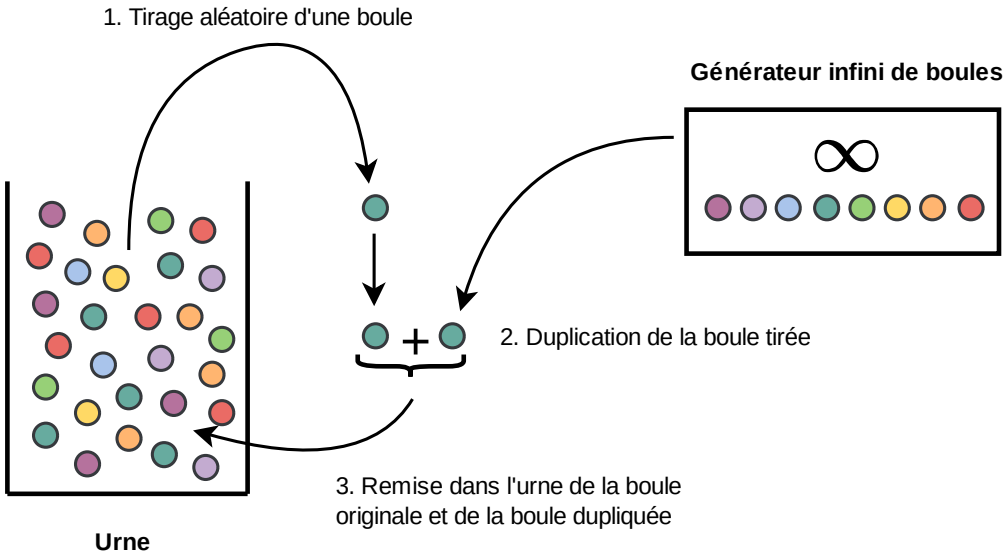


FIGURE 5.2 – Tirage et remise pour une urne de Pólya simple. À chaque tirage d'une boule de l'urne, sa couleur est observée et une boule de la même couleur est générée. La boule tirée et la boule dupliquée ces deux boules sont ajoutées à l'urne.

la probabilité de tirer une boule de couleur u' de l'urne v est égale au quotient du nombre de boules de couleur u' dans v (boules ajoutées par sur-remise $n_{vu'}^{-(uo)}$ et boules initiales $\mu\frac{1}{U}$) par le nombre de boules total dans v (boules ajoutées par sur-remise $n_{v\bullet}^{-(uo)}$ et boules initiales μ). De plus, le terme le plus à droite dans l'Équation (5.4) est exactement le même que celui de l'Équation (5.6) obtenu en suivant SPU, illustrant ainsi l'équivalence entre la distribution composée (*compound distribution*) Dirichlet-multinomiale et le processus de SPU.

Notons que nous avons précédemment supposé que la boule (uo) était la dernière à être tirée. Cette hypothèse est applicable en raison de la nature échangeable (*exchangeable*) de SPU : toute permutation d'une séquence de tirages de boules donnée est associée à une même probabilité jointe. En d'autres termes, l'ordre dans lequel les couleurs des boules tirées sont observées n'influence pas la probabilité du dernier tirage. Ce résultat est associé au théorème de de Finetti – le lecteur pourra se référer à [Aldous, 1985] pour plus de détails sur le sujet.

5.2.2 Urnes de Pólya généralisées

Le processus d'urne de Pólya généralisée (GPU – *generalized Pólya urn*) [Mahmoud, 2008] étend SPU en changeant la règle de remise : suite au tirage d'une boule de couleur c, la boule est remise dans l'urne et un certain nombre de boules (ou parties de boules) de couleurs « similaires » à c sont également ajoutés à l'urne. Dans les modèles thématiques précédents utilisant GPU [Chen et Liu, 2014; Li *et al.*, 2016; Mimno *et al.*, 2011; Wang *et al.*, 2016], les boules correspondent aux emplacements de mots (*tokens*) apparaissant dans un document et les couleurs des boules sont les mots du vocabulaire associés à ces emplacements. Dans ce cas, la similarité entre couleurs, utilisée pour choisir les différentes boules à remettre dans

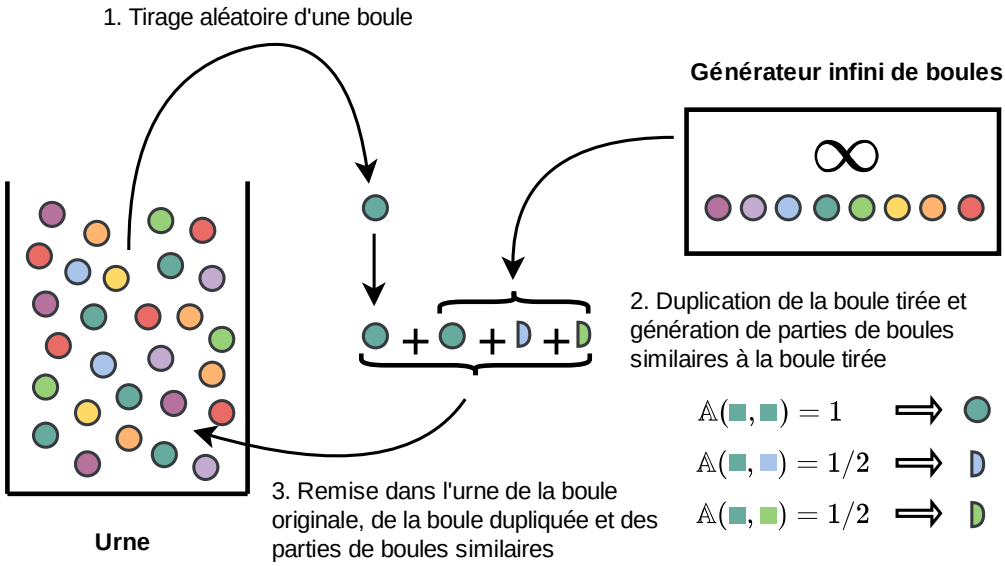


FIGURE 5.3 – Tirage et remise pour une urne de Pólya généralisée. À la différence de l'urne de Pólya simple, on ajoute également des portions de boules de couleurs similaires à la couleur de la boule tirée. Pour une couleur donnée tirée, la proportion de boule de chaque couleur à ajouter à l'urne est fournie par la matrice $\mathbb{A}$. Dans cet exemple, les valeurs de $\mathbb{A}$ que nous n'avons pas spécifiées sont considérées comme étant nulles.

l'urne, est sémantique et par exemple basée sur la PMI [Chen et Liu, 2014; Wang *et al.*, 2016] ou bien calculée à partir de plongements de mots (*word embeddings*) [Li *et al.*, 2016]. L'avantage d'utiliser GPU dans ce contexte est de tenir compte des dépendances entre les mots lors de leur attribution à un thème : si les mots sémantiquement similaires a un mot w sont majoritairement attribués à un thème z, alors il semble raisonnable de promouvoir le thème z pour w.

Dans le reste de ce chapitre, nous dénoterons par $\mathbb{A}$ la matrice à valeurs réelles comprises entre 0 et 1 indiquant la portion de boule à ajouter à l'urne pour chaque couleur, sachant la couleur de la boule tirée. $\mathbb{A}$ est connue dans la littérature sous le nom de *matrice d'addition* (*addition matrix*) ou de *schéma* (*schema*) [Mimno *et al.*, 2011]. Par exemple, si une boule de couleur c est tirée de l'urne, on ajoutera à l'urne $\mathbb{A}_{cc'}$ boule pour chaque couleur $c' = 1, ..., C$. En pratique, $\mathbb{A}$ est une matrice creuse et chaque ligne contient seulement un nombre limité d'éléments non nuls – c'est-à-dire un nombre limité de couleurs similaires à la couleur associé à la ligne. La procédure complète de tirage et de remise pour GPU est illustrée dans la Figure 5.3.

5.2.3 Description de SNVDM-GPU

À la différence des modèles précédents qui intègrent GPU pour exploiter la similarité sémantiques des mots dans les documents, nous proposons dans notre extension SNVDM-GPU d'utiliser le processus de GPU pour prendre en compte les liens faibles existant entre

les utilisateurs d'un réseau social. Comme dans l'exemple que nous avons présenté dans la Section 5.2.1, nous considérons les interactions comme étant des boules et les accointances associées à ces interactions comme étant des couleurs. En s'appuyant sur la topologie du réseau, nous supposons que les utilisateurs les plus similaires à un utilisateur u sont les utilisateurs qui interagissent le plus avec u. Cette hypothèse découle du principe d'homophilie.

Définissons tout d'abord formellement l'ensemble des utilisateurs similaires à un utilisateur u donné. Le nombre d'interactions entrantes sur u où u' est l'utilisateur source comme $i_{uu'} = \sum_{d=1}^{D_u} \sum_{i=1}^{I_{ud}} \mathbb{I}(s_{udi} = u')$ et le nombre d'interactions sortantes de u où u' est l'utilisateur cible comme $o_{uu'} = \sum_{o=1}^{O_u} \mathbb{I}(r_{uo} = u')$. Comme précédemment, $\mathbb{I}$ désigne la fonction indicatrice : $\mathbb{I}(\text{vrai}) = 1$ et $\mathbb{I}(\text{faux}) = 0$. L'ensemble des τ accointances de l'utilisateur u interagissant le plus u (si un tel nombre d'accointances existe pour u), noté $\mathcal{R}_{u\tau}$, possède alors les trois propriétés suivantes :

$$|\mathcal{R}_{u\tau}| = \min\left(\tau,\ \sum_{u'=1}^{U} \mathbb{I}(i_{uu'} + o_{uu'} > 0)\right); \tag{5.7}$$

$$\forall u' \in \mathcal{R}_{u\tau},\ i_{uu'} + o_{uu'} > 0\,; \tag{5.8}$$

$$\forall u' \in \mathcal{R}_{u\tau},\ \forall u'' \notin \mathcal{R}_{u\tau},\ i_{uu'} + o_{uu'} \geq i_{uu''} + o_{uu''}. \tag{5.9}$$

La première propriété (Équation (5.7)) indique simplement que la taille de $\mathcal{R}_{u\tau}$ ne peut être supérieure au nombre total d'utilisateurs qui interagissent avec u. La deuxième propriété (Équation (5.8)) exprime le fait que tout utilisateur u' de $\mathcal{R}_{u\tau}$ est impliqué dans au moins une interaction (entrante ou sortante) avec u – c'est-à-dire que u' est une accointance de u. La troisième propriété (Équation (5.9)) garantit que les utilisateurs de $\mathcal{R}_{u\tau}$ interagissent plus avec u que ne le font les utilisateurs hors de $\mathcal{R}_{u\tau}$.

À partir de $\mathcal{R}_{u\tau}$, nous établissons maintenant la définition de la matrice d'addition $\mathbb{A}$ dans le cas de l'extension de SNVDM. Pour une boule de couleur u tirée d'une urne de type GPU, $\mathbb{A}_{uu'}$ exprime la portion de boule de couleur u' ajoutée à l'urne lors de la remise, pour chaque $u' = 1, ..., U$:

$$\mathbb{A}_{uu'} = \begin{cases} 1 & \text{if } u = u', \\ \lambda & \text{if } u \neq u' \text{ and } u' \in \mathcal{R}_{u\tau}, \\ 0 & \text{otherwise}\,; \end{cases} \tag{5.10}$$

où λ est un paramètre à valeur réelle dans l'intervalle $[0\,;1]$ dénotant les portions de boule à ajouter pour les couleurs similaires à u (c'est-à-dire les couleurs dans $\mathcal{R}_{u\tau}$).

Contrairement à SPU, une séquence de boules tirées suivant GPU n'est pas échangeable : l'ordre dans lequel les boules sont tirées a un impact sur la probabilité jointe des couleurs de la séquence de boules. Cette propriété rend le calcul des probabilités conditionnelles postérieures – destinées à être utilisées dans l'échantillonneur de Gibbs marginalisé de notre extension SNVDM-GPU – beaucoup plus coûteux. Par conséquent, suivant [Mimno *et al.*, 2011], nous donnons une approximation de la probabilité conditionnelle postérieure exacte en supposant que la boule ou la séquence de boules d'intérêt est tirée en dernier, comme si l'on

avait déjà tiré et observé les couleurs de toutes les autres boules (interactions) de la collection. Nous ignorons donc les implications du tirage courant sur les futurs tirages de boules.

L'intégration du processus de GPU dans SNVDM – aboutissant au modèle SNVDM-GPU – sous cette approximation mène à une légère modification de l'échantillonneur de Gibbs marginalisé décrit dans la Section 5.1.3. Lors de l'échantillonnage du point de vue d'un document v_{ud}, on peut montrer que la probabilité $p(\mathbf{s}_{ud}|\mathbf{s}^{-(ud)}, v_{ud} = v, \mathbf{v}^{-(ud)})$, correspondant au deuxième terme de l'Équation (5.3) :

$$\frac{\prod_{u'=1}^{U}\prod_{a=0}^{n_{vu'}^{(ud)}-1}(n_{vu'}^{-(ud)} + a + \mu\frac{1}{U})}{\prod_{b=0}^{n_{v\bullet}^{(ud)}-1}(n_{v\bullet}^{-(ud)} + b + \mu)} \tag{5.11}$$

devient la suivante en appliquant GPU et l'approximation sus-mentionnée :

$$\prod_{i=1}^{I_{ud}} \frac{\sum_{u''=1}^{U}\mathbb{A}_{u''s_{udi}}\, n_{vu''}^{-(ud)} + \sum_{j=1}^{i-1}\mathbb{A}_{s_{udj}s_{udi}} + \mu\frac{1}{U}}{\sum_{u''=1}^{U}\mathbb{A}_{u''\bullet}\, n_{vu''}^{-(ud)} + \sum_{j=1}^{i-1}\mathbb{A}_{s_{udj}\bullet} + \mu} \tag{5.12}$$

dans laquelle l'ordre de tirage utilisée pour la séquence $\mathbf{s}_{ud}$ est l'ordre arbitraire fourni par l'indice $i = 1, ..., I_{ud}$.

Similairement, la probabilité $p(r_{uo} = u'|r^{-(uo)}, v'_{uo} = v, \mathbf{v'}^{-(uo)})$, correspondant au second terme de l'Équation (5.4) pour le tirage du point de vue d'une interaction sortante v'_{uo} :

$$\frac{n_{vu'}^{-(uo)} + \mu\frac{1}{U}}{n_{v\bullet}^{-(uo)} + \mu} \tag{5.13}$$

doit être remplacé dans SNVDM-GPU par le terme suivant :

$$\frac{\sum_{u''=1}^{U}\mathbb{A}_{u''u'}\, n_{vu''}^{-(uo)} + \mu\frac{1}{U}}{\sum_{u''=1}^{U}\mathbb{A}_{u''\bullet}\, n_{vu''}^{-(uo)} + \mu}. \tag{5.14}$$

5.3 Expérimentations

Dans cette section, nous détaillons les expérimentations que nous avons conduites afin d'évaluer nos modèles SNVDM et SNVDM-GPU, et leur comparaison aux approches de l'état de l'art. Nous avons pour objectif de tester les hypothèses suivantes :

— **(H1)** La performance en terme de regroupement de points de vue des modèles que nous proposons **(a)** surpasse celle des approches de l'état de l'art ; elle est améliorée par **(b)** l'exploitation des interactions entrantes et sortantes et **(c)** l'intégration du processus d'urne de Pólya généralisée.

— **(H2)** Le processus d'urne de Pólya généralisée est bénéfique à la robustesse des modèles proposés dans le cas où le réseau social est peu dense (*social network sparsity*) – c'est-

à-dire que les modèles basés sur GPU sont moins affectés par la limitation du nombre d'interactions sociales.

— **(H3)** Les modèles proposés ont un temps d'exécution comparable (même ordre de magnitude) à celui des modèles de l'état de l'art.

— **(H4)** Les points de vue et les thèmes découverts par notre approche sont cohérents.

Dans le reste de cette section, nous introduisons dans un premier temps notre cadre expérimental (Section 5.3.1). Nous détaillons ensuite les différentes évaluations quantitatives et qualitatives conduites sur nos modèles et approches de l'état de l'art (Section 5.3.2).

5.3.1 Cadre expérimental

Nous explicitons le cadre expérimental dans cette section en présentant les modèles de référence adoptés pour la comparaison avec SNVDM et SNVDM-GPU (Section 5.3.1.1), les collections sur lesquelles ont porté les expérimentations (Section 5.3.1.2) et les paramètres que nous avons choisis (Section 5.3.1.3).

5.3.1.1 Modèles de référence

Afin de tester les hypothèses (H1)-(H4), nous comparons les modèles que nous proposons à différentes approches existantes capables de découvrir les points de vue ou des dimensions latentes similaires :

— **TAM** (*Topic Aspect Model*) [Paul et Girju, 2010] a été initialement défini pour découvrir conjointement les thèmes et les aspects – qui représentent les points de vue dans notre contexte. Ce modèle a été précédemment décrit dans les Section 2.3.1.1 et 4.2.1.1. TAM ne prend pas en compte les interactions entre utilisateurs.

— **SN-LDA** (*Social Network Latent Dirichlet Allocation*) [Sachan *et al.*, 2014] a pour objectif de découvrir conjointement les thèmes et les communautés – que nous interprétons ici comme des points de vue – dans un réseau social où les utilisateurs sont associés à un contenu textuel. Les seules interactions utilisées par SN-LDA sont les interactions sortantes.

— **VODUM** (*Viewpoint and Opinion Discovery Unification Model*) [Thonet *et al.*, 2016] est le modèle que nous avons introduit dans le Chapitre 4. Il modélise conjointement les points de vue et les thèmes, et exploite les parties de discours pour mieux différencier les mots d'opinions (c'est-à-dire les mots de point de vue thématiques) des mots thématiques. VODUM n'utilise pas les interactions entre utilisateurs.

Les approches dérivant des modèles que nous proposons dans ce chapitre que nous testons ici sont les suivantes :

— **SNVDM** est le premier modèle que nous proposons, décrit dans la Section 5.1. Il exploite à la fois les interactions entrantes et sortantes.

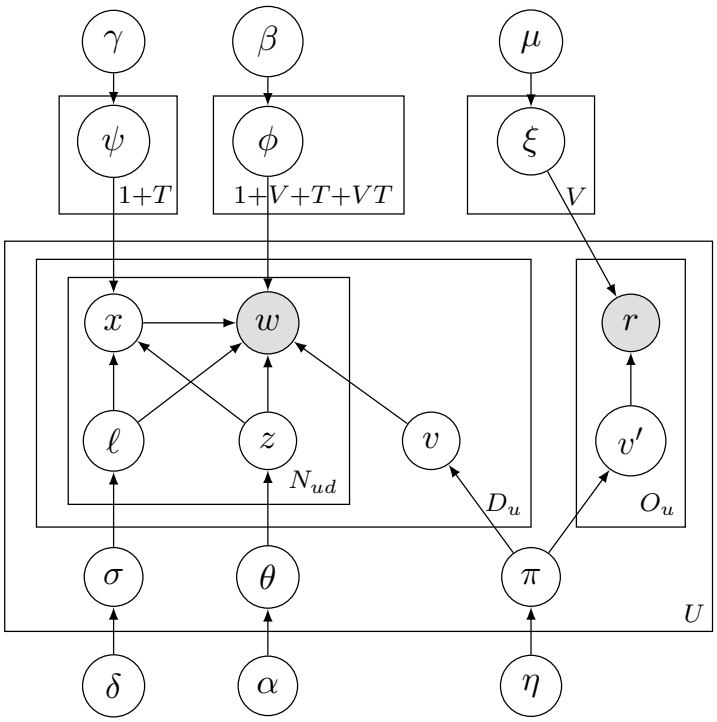


FIGURE 5.4 – Représentation sous forme de modèle graphique de SNVDM-WII. Comparé à SNVDM, SNVDM-WII n'utilise pas les interactions entrantes, d'où l'absence des variables observées **s** (utilisateurs sources des interactions entrantes).

— **SNVDM-WII** est une version dégénérée de SNVDM qui n'utilise pas les interactions entrantes. Les seules interactions utilisées sont les interactions sortantes. Nous testons ce modèle pour vérifier si l'utilisation des interactions entrantes est bénéfique à SNVDM.

— **SNVDM-GPU ($\tau = 10$)** et **SNVDM-GPU ($\tau = \infty$)** étendent SNVDM en intégrant le processus d'urne de Pólya généralisée lors de l'inférence postérieure approchée par échantillonnage de Gibbs marginalisé, comme décrit dans la Section 5.2. La valeur de τ définit le nombre maximum d'accointances interagissant le plus avec un utilisateur donné à intégrer en tant qu'utilisateurs similaires à cet utilisateur dans la matrice $\mathbb{A}$. Dans SNVDM-GPU ($\tau = 10$), seules les 10 accointances de chaque utilisateur qui possèdent le plus d'interactions sont utilisées, tandis que pour SNVDM-GPU ($\tau = \infty$) toutes les accointances d'un utilisateur donné sont considérées comme étant les utilisateurs qui lui sont similaires.

Le code Java des modèles proposés et des modèles de référence, utilisés dans nos expérimentations, est accessible sur `https://github.com/tthonet/SNVDM`.

5.3.1.2 Collections de données

Il existe un nombre très limité de collections permettant l'évaluation de modèles pour la découverte de points de vue sur les réseaux sociaux. En 2016, SemEval (*International Work-*

shop on Semantic Evaluation) a introduit une tâche pour la détection de positions (*stance detection*) sur Twitter [Mohammad *et al.*, 2016]. Cependant, cette tâche se focalise sur la détection de position au niveau des *tweets*, tandis que dans ce chapitre nous sommes interessé par la découverte de points de vue au niveau des utilisateurs. Les méthodes et les données pertinentes à ces deux problèmes sont distinctes étant donné que seul le second permet l'exploitation des interactions sociales. Par conséquent, nous avons choisi de valider nos approches sur deux collections Twitter [3] introduites dans [Brigadir *et al.*, 2015], auquel nous référons le lecteur pour des détails complémentaires sur la construction des collections. La première collection, que nous dénoterons ci-après par *Indyref*, contient des *tweets* concernant le référendum sur l'indépendance de l'Écosse en 2014, postés entre 11/08/2014 et 20/10/2014. Les deux points de vue représentés sont le *Oui* et le *Non*, exprimant respectivement le soutien et l'opposition à l'indépendance de l'Écosse. La seconde collection, que nous dénoterons ci-après par *Midterms*, est constituée de *tweets* rédigés par les acteurs politiques actifs durant les élections de mi-mandat (*midterm elections*) aux États-Unis en 2014. Notons que pour étendre la collection Midterms nous avons utilisé l'historique (*timeline*) Twitter complet des utilisateurs (jusqu'au 21/11/2014) au lieu de nous restreindre aux *tweets* postés durant la période des élections de mi-mandat. Seuls les utilisateurs *démocrates* et *républicains* apparaissent dans cette collection. Pour chaque collection, la vérité terrain sur le point de vue des utilisateurs a été obtenue dans [Brigadir *et al.*, 2015] soit en utilisant les utilisateurs indiquant explicitement leur orientation politique dans leurs profils, soit à partir de listes Twitter officielles et non-officielles. Similairement aux travaux précédents sur la polarisation politique [Conover *et al.*, 2011b; Magdy *et al.*, 2016], les interactions sociales de Twitter que nous avons considérées pour définir le réseau social reliant les utilisateurs sont les *retweets* et les *replies*.

Un problème auquel nous avons dû faire face vis-à-vis des collections est le fait que celles-ci contiennent un grand nombre de *tweets* qui ne concernent pas le sujet politique étudié – par exemple, des *tweets* sur les loisirs ou le quotidien. Afin d'éliminer ces *tweets* non pertinents au sujet d'étude, nous avons ignoré tous les *tweets* sur lesquels les utilisateurs du dataset n'ont pas interagi, c'est-à-dire les tweets sans réponses ni *retweets*. Bien que ce processus mène à la suppression d'un certain nombre de *tweets* pertinents, nous avons observé que cela nous permet d'obtenir des collections considérablement plus focalisées sur le sujet. De plus, nous avons seulement conservé les *tweets* uniques dans les collections (en éliminant les duplicata associés aux *retweets*) et nous les avons assignés à leurs auteurs originaux. Après l'étape d'élimination des *tweets* non pertinents, les utilisateurs sans *tweets* associés ont été éliminés (32 pour Indyref et 232 pour Midterms). Notons que ce processus est applicable à toute collection contenant des interactions sociales et ne nécessite pas de supervision. Cependant, cette opération affecte le ratio du nombre d'interactions par *tweet* ; par conséquent, nous étudions plus en détail dans la Section 5.3.2 l'impact d'un réseau social de faible densité sur les performances de nos modèles et celles des approches de référence.

Nous avons ensuite effectué les étapes suivantes de prétraitement sur les collections en utilisant Lingpipe [4] et TweetNLP [5]. Nous avons appliqué aux *tweets* un étiqueteur de partie

3. http://dx.doi.org/10.6084/m9.figshare.1430449
4. http://alias-i.com/lingpipe/
5. http://www.cs.cmu.edu/~ark/TweetNLP/

Tableau 5.2 – Statistiques des collections utilisées dans nos expérimentations. Les interactions rapportées comprennent à la fois les interactions entrantes et sortantes. « Dém. » et « Rép. » indiquent respectivement les points de vue démocrate et républicain.

Collection	#Utilisateurs		# *Tweets*	#Mots	Vocabulaire	#Interactions
	Oui/Dém.	Non/Rép.				
Indyref	589	575	270 075	2 043 204	38 942	696 654
Midterms	767	778	113 545	975 199	25 312	241 741

du discours fourni par TweetNLP, étant donné que cela est nécessaire pour le modèle de référence VODUM afin de distinguer les mots d'opinion et les mots thématiques. Suivant la procédure décrite dans la Section 4.2.1.2, les noms ont été considérés comme mots thématiques et les verbes, adverbes, adjectifs et prépositions (et également les *hastags*, qui peuvent être utilisés de manière partisane) ont été utilisés en tant que mots d'opinion dans VODUM. Les mots attribués à des parties de discours autres que celles évoquées précédemment ont été éliminés des collections utilisées par tous les modèles – en effet, pour effectuer une comparaison équitable, nous souhaitions que les modèles soient testés exactement sur les mêmes données textuelles. Enfin, nous avons enlevé les mots vides, les noms d'utilisateurs, les URLs et les mots n'apparaissant qu'une fois. Suivant [Schofield et Mimno, 2016], nous n'avons pas effectué de racinisation des collections. Les utilisateurs sans *tweets* (ou avec des *tweets* pour lesquels il ne reste aucun mot) ont été éliminés. Les statistiques des collections après les différentes étapes de prétraitement sont détaillées dans le Tableau 5.2.

5.3.1.3 Choix des paramètres

Pour les deux collections Indyref et Midterms, nous avons fixé à 2 le nombre de points de vue (pour VODUM et pour les modèles basés SNVDM et SNVDM-GPU), le nombre d'aspects (pour TAM) et le nombre de communautés (pour SN-LDA). Les hyperparamètres de chaque modèle ont été initialisés à 1 puis mis à jour suivant la procédure décrite dans la Section 5.1.3, à l'exception des hyperparamètres des distributions de mots (β pour les modèles basés sur SNVDM et SNVDM-GPU) qui ont été fixés à 0,01. Le paramètre λ utilisé dans les modèles intégrant GPU – spécifiant la portion de boule à remettre à l'urne pour les accointances d'accointances – a été fixé à 0,5. Ce choix a été guidé par l'intuition selon laquelle l'influence d'une accointance d'accointance (remise d'une demi boule à l'urne) devrait être moins importante que l'influence d'une accointance directe (remise d'une boule entière). Des expériences préliminaires que nous avons conduites ont par ailleurs confirmé que l'utilisation de $\lambda = 0{,}5$ mène à de meilleures résultats que $\lambda = 1$ (équivalant à un traitement uniforme des accointances directes et des accointances d'accointances).

L'inférence de l'échantillonneur de Gibbs marginalisé des modèles proposés et modèles de référence a été effectuée sur 5 chaînes de Markov différentes (c'est-à-dire 5 exécutions à partir d'initialisations aléatoires différentes) pour 1 000 itérations chacune, avec 500 itérations de

burn-in. Après le *burn-in*, un tirage est conservé toutes les 50 itérations et les distributions des modèles sont estimées en se basant sur les 10 tirages collectés.

5.3.2 Évaluation

Dans cette section, nous décrivons les résultats obtenus par les modèles de référence et les modèles proposés à travers différentes expérimentations sur les collections Indyref et Midterms. Nous rapportons tout d'abord les performances des différents modèles en terme de regroupement de points de vue (*viewpoint clustering*) dans la Section 5.3.2.1. La Section 5.3.2.2 étudie ensuite la robustesse des modèles exploitant les interactions entre utilisateurs lorsque le réseau social est peu dense (*sparse*). Nous discutons également de l'efficacité de notre modèle en terme de temps d'exécution (*efficiency*) dans la Section 5.3.2.3. Enfin, la Section 5.3.2.4 fournit une analyse qualitative de notre approche en montrant plusieurs thèmes et points de vue découverts.

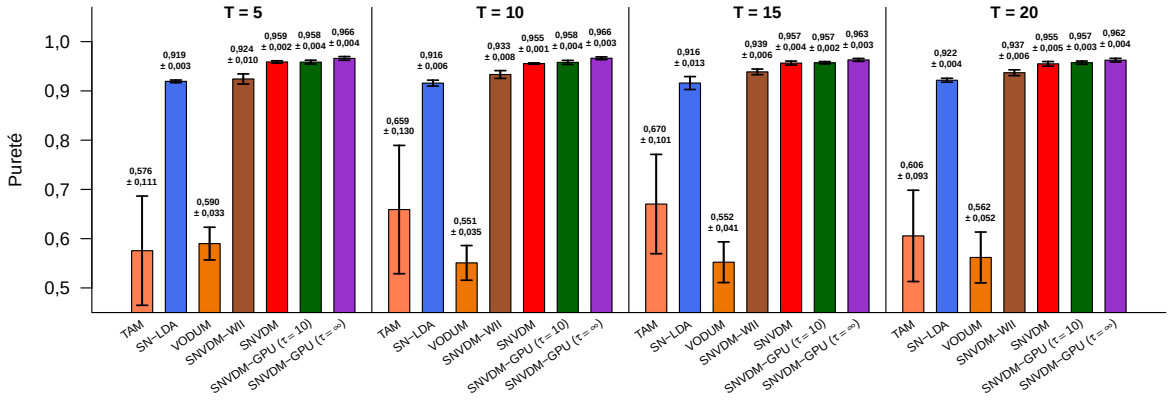
5.3.2.1 Regroupement de points de vue

Nous avons comparé nos modèles SNVDM, sa version dégénérée SNVDM-WII, ainsi que SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) aux modèles de référence TAM, SN-LDA et VODUM sur une tâche de regroupement d'utilisateurs en fonction de leur point de vue. Les métriques que nous avons adoptées pour rapporter les performances des différents modèles sont la Pureté (*Purity*) et l'Information Mutuelle Normalisée (NMI – Normalized Mutual Information) [Manning *et al.*, 2008]. La Pureté mesure la proportion d'utilisateurs qui sont assignés à leur classe correcte fournie par la vérité terrain. La NMI est une mesure de regroupement issue de la théorie de l'information et basée sur l'Information mutuelle et l'entropie. Nous avons également testé la mesure BCubed F [Amigó *et al.*, 2009] mais nous avons obtenu une corrélation de rang de Spearman (*Spearman rank correlation*) presque parfaite entre la Pureté et BCubed F ($\rho = 0,998$) et entre la NMI et BCubed F ($\rho = 0,999$) – calculée à partir des 280 mesures obtenus dans le reste de cette section, sur les différentes collections et différents nombres de thèmes. Par conséquent, nous rapportons ici les résultats de regroupement seulement selon les mesures plus classiques de Pureté et de NMI.

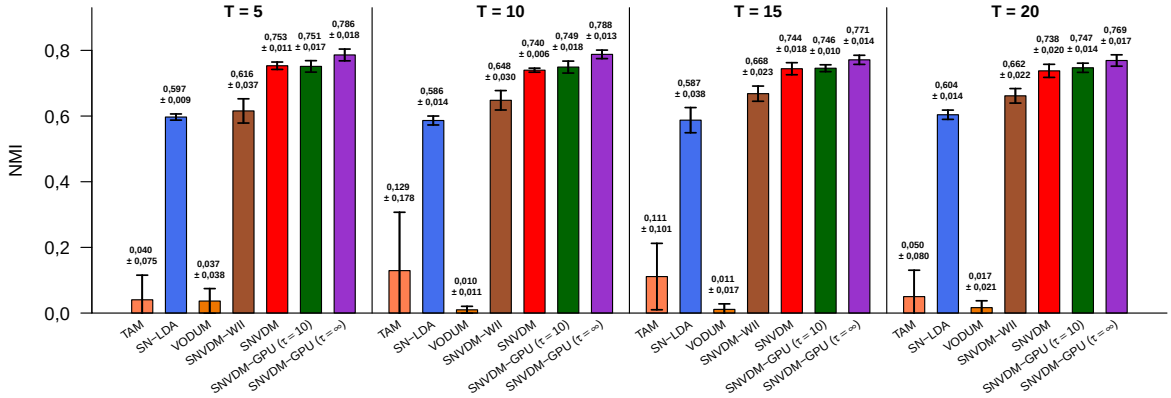
Les groupes de points de vue (*viewpoint clusters*) sont construits comme suit. Dans les modèles basés sur SNVDM et SNVDM-GPU, chaque utilsateur u se voit assigner au point de vue (groupe) qui maximise la distribution de points de vue $\{\pi_{uv}\}_{v=1}^{V}$ spécifique à l'utilisateur u. Les attributions d'aspect (pour TAM) et de communauté (pour SN-LDA) aux utilisateurs sont obtenus de manière similaire. Dans VODUM, un point de vue est naturellement attribué à chaque utilisateur, étant donné que les points de vue sont définis au niveau de l'utilisateur [6].

Les résultats du regroupement de points de vue en termes de Pureté et de NMI sur les collections Indyref et Midterms sont rapportés dans les Figures 5.5 (Indyref) et 5.6 (Midterms)

6. Dans ce chapitre, chaque utilisateur, agrégat de l'ensemble des *tweets* rédigé par celui-ci, est considéré comme un document pour TAM et VODUM.



(a) Puretés obtenues sur la collection Indyref pour différents nombres de thèmes.

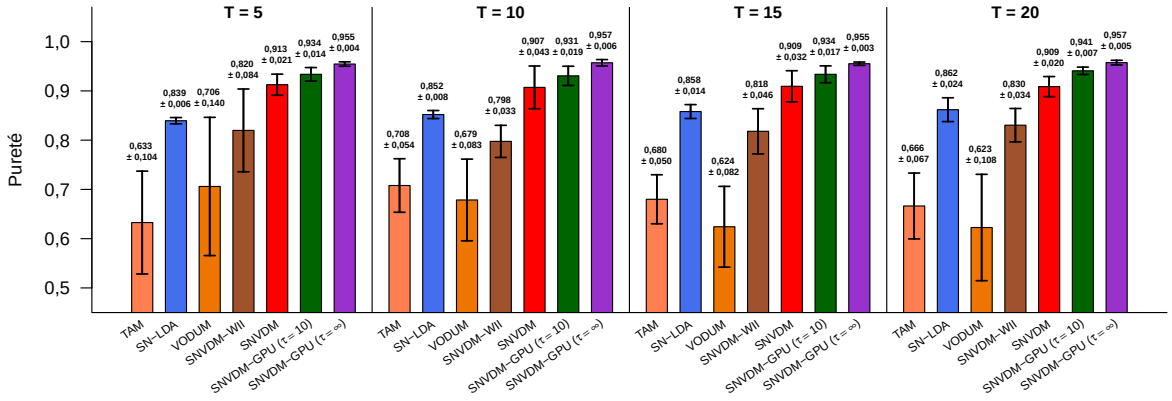


(b) NMI obtenues sur la collection Indyref pour différents nombres de thèmes.

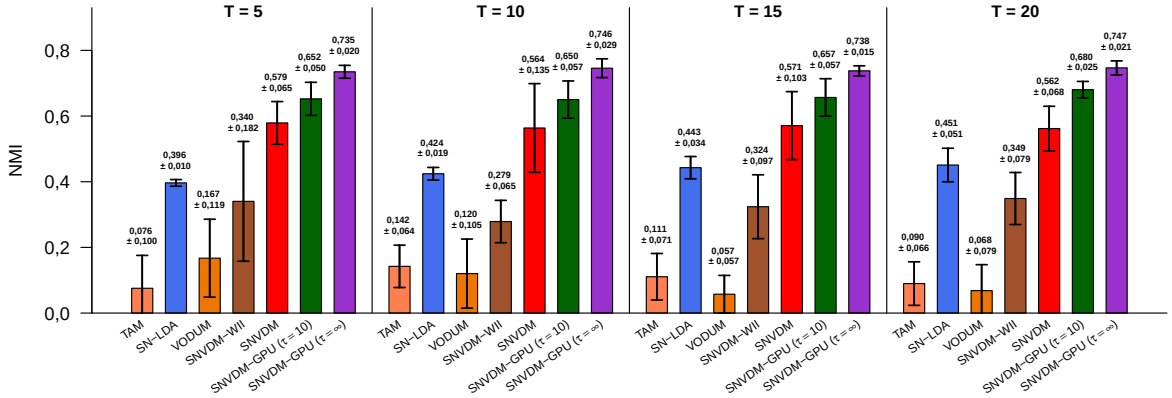
FIGURE 5.5 – Résultats du regroupement de points de vue sur la collection **Indyref** en terme de Pureté (5.5a) et de NMI (5.5b) pour les modèles TAM, SN-LDA, VODUM, SNDM-WII, SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) avec différents nombres de thèmes (de gauche à droite : 5, 10, 15 et 20). Une valeur plus élevée indique une meilleure performance. Les barres d'erreur dénotent les intervalles de confiance à 95 % autour de la moyenne, calculés à partir de 5 exécutions suivant une loi t de Student.

pour différents nombres de thèmes $T \in \{5, 10, 15, 20\}$. Les barres d'erreur dénotent des intervalles de confiance à 95 % autour de la moyenne, calculés à partir des 5 exécutions répétées (c'est à dire des 5 chaînes de Markov). En confirmation de l'hypothèse (H1a), nous observons que sur les deux collections nos modèles SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) surpassent les modèles de référence avec une Pureté moyenne et une NMI moyenne supérieures. Le fait que leurs intervalles de confiance à 95 % ne se superposent pas à ceux des modèles de référence confirme également la significativité statistique de cette observation à un niveau de significativité $\alpha = 0{,}05$. Nous notons également que nos modèles semblent essentiellement insensibles au nombre de thèmes utilisé.

La comparaison de nos modèles SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) avec SN-LDA et la version dégénérée SNVDM-WII, qui n'exploitent pas les in-



(a) Puretés obtenues sur la collection Midterms pour différents nombres de thèmes.



(b) NMI obtenues sur la collection Midterms pour différents nombres de thèmes.

FIGURE 5.6 – Résultats du regroupement de points de vue sur la collection **Midterms** en terme de Pureté (5.6a) et de NMI (5.6b) pour les modèles TAM, SN-LDA, VODUM, SNDM-WII, SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) avec différents nombres de thèmes (de gauche à droite : 5, 10, 15 et 20). Une valeur plus élevée indique une meilleure performance. Les barres d'erreur dénotent les intervalles de confiance à 95 % autour de la moyenne, calculés à partir de 5 exécutions suivant une loi t de Student.

teractions entrantes, soutient l'hypothèse (H1b). En effet, tous les modèles qui utilisent les interactions entrantes en plus des interactions sortantes surpassent significativement celles qui n'utilisent que les interactions sortantes. Nous notons néanmoins que le modèle de référence SN-LDA a également obtenu de bons résultats de regroupement sur les deux collections, et en particulier sur Indyref. Cela peut être expliqué par le fait que Indyref présente un grand nombre d'interactions par utilisateur, desquelles SN-LDA dépend fortement. Cela souligne également l'importance clé des interactions sociales pour la découverte de points de vue dans les cas étudiés. Les autres modèles de référence TAM et VODUM, qui ne font pas usage des interactions, ont obtenu des résultats nettement inférieurs à ceux de SN-LDA. Globalement, TAM et VODUM ont atteint des performances comparables sur Midterms et TAM a obtenu de meilleurs résultats sur Indyref. L'infériorité de VODUM comparé à TAM dans ce cadre pourrait être due au fait que VODUM repose sur les parties de discours. Celles-ci sont poten-

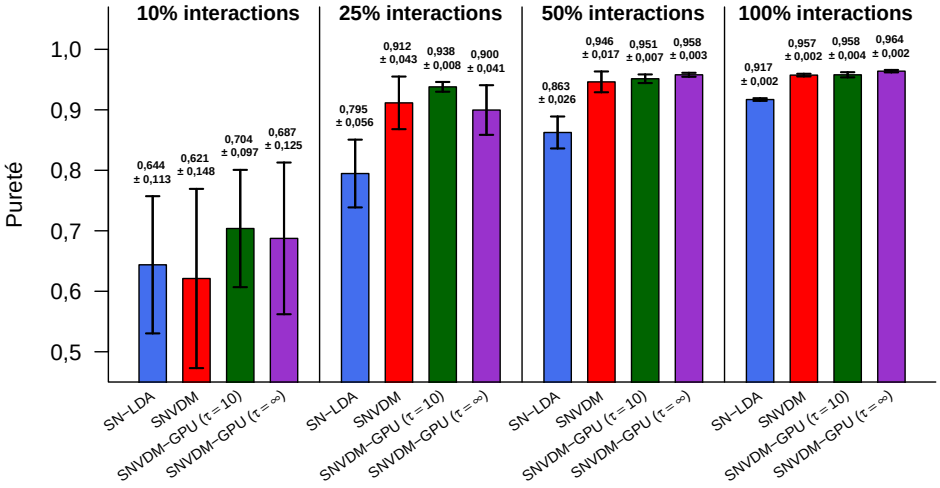
tiellement moins discriminantes dans l'identification de mots d'opinion dans des textes bruités et courts tels que des *tweets* qu'elles ne le sont pour des textes longs et formels, comme nous l'avons vu dans le Chapitre 4. VODUM et TAM semblent modérément sensibles au nombre de thèmes, VODUM obtenant de meilleurs résultats pour $T = 5$ sur les deux collections.

L'hypothèse (H1c) est également validée en observant que les modèles SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$), qui intègrent le processus GPU, ont atteint des performances de regroupement légèrement meilleures que celle de SNVDM. Toutefois, notons que SNVDM-GPU ($\tau = 10$) surpasse significativement SNVDM à un niveau $\alpha = 0{,}05$ seulement sur la collection Midterms pour $T = 20$. Sur Indyref, leurs intervalles de confiance à 95 % se superposent à la fois pour la Pureté et la NMI ; la différence n'est donc pas significative. Nous soupçonnons le grand nombre d'interactions de la collection Indyref d'être responsable des performances similaires de SN-LDA, SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$). Par conséquent, nous étudions dans la Section 5.3.2.2 dans quelle mesure ces différents modèles sont robustes aux réseaux sociaux de plus faible densité, c'est-à-dire lorsque les interactions entre utilisateurs sont moins nombreuses.

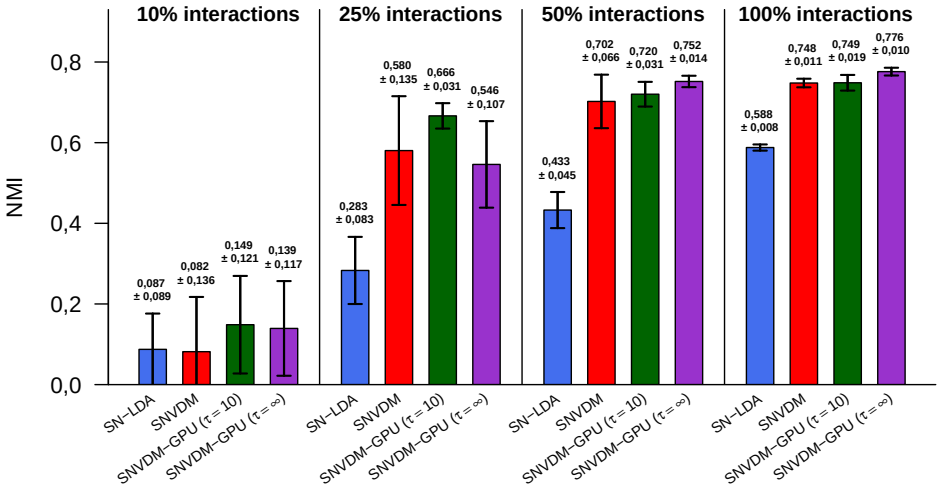
5.3.2.2 Robustesse aux réseaux sociaux de faible densité

Afin d'étudier la robustesse des modèles dans le cas où le réseau social des utilisateurs est peu dense, nous avons artificiellement conservé différents pourcentages (100 %, 50 %, 25 % et 10 %) du nombre total d'interactions disponibles dans la collection Indyref. Les interactions à retirer – entrantes ou sortantes – ont été choisies aléatoirement. Pour cela, nous avons effectué pour chaque interaction un tirage suivant une loi de Bernoulli paramétrisée en fonction du pourcentage d'interactions désirées. Par exemple, si le nombre d'interactions souhaitées est 25 % du total, la probabilité de succès est fixée à 1/4. Une fois ce traitement appliqué à Indyref, nous analysons la performance de SN-LDA, SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) pour les différents pourcentages d'interactions retenues. Étant donné que le nombre de thèmes a seulement un impact mineur sur la performances des différents modèles – comme observé dans la Section 5.3.2.1, nous avons fixé $T = 10$ dans cette expérimentation. Les résultats sont présentés dans la Figure 5.7. Ici également, nous indiquons les intervalles de confiance à 95 % autour de la moyenne. Nous observons une tendance similaire pour tous les modèles : la performance de regroupement de points de vue est dégradée de manière substantielle pour un faible pourcentage d'interactions, et en particulier pour 10 %. Ce résultat souligne de nouveau que les interactions sont clés pour l'identification des points de vue dans un réseau social.

Par ailleurs, dans le cas étudié, les modèles basés sur GPU semblent seulement marginalement plus robustes aux réseaux peu denses que les autres modèles, avec une amélioration significative uniquement par rapport à SN-LDA pour 100%, 50% et 25% des interactions. Cette observation n'apporte donc qu'un soutien nuancé à l'hypothèse (H2). Cependant, il est intéressant de constater que SNVDM-GPU ($\tau = 10$) paraît légèrement plus robuste aux réseaux peu denses que SNVDM-GPU ($\tau = \infty$) : en dessous de 50 %, SNVDM-GPU ($\tau = 10$) a obtenu de meilleures performances. Cela pourrait être expliqué par le fait que SNVDM-GPU



(a) Puretés obtenues sur la collection Indyref pour différents pourcentages des interactions disponibles.



(b) NMI obtenues sur la collection Indyref pour différents pourcentages des interactions disponibles.

FIGURE 5.7 – Résultats du regroupement de points de vue sur la collection **Indyref** en terme de Pureté (5.7a) et de NMI (5.7b) pour les modèles SN-LDA, SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) en conservant différents pourcentages des interactions disponibles dans le réseau social (de gauche à droite : 10 %, 25 %, 50 % et 100 %). Le nombre de thèmes a été fixé à 10. Une valeur plus élevée indique une meilleure performance. Les barres d'erreur dénotent les intervalles de confiance à 95 % autour de la moyenne, calculés à partir de 5 exécutions suivant une loi t de Student.

($\tau = 10$) est plus « sélectif » que SNVDM-GPU ($\tau = \infty$) dans le choix des accointances à inclure dans $\mathbb{A}$. En effet, le premier modèle utilise les 10 accointances de chaque utilisateur possédant le plus d'interactions avec celui-ci – qui sont possiblement plus représentatives du point de vue de l'utilisateur – tandis que le second exploite toutes les accointances disponibles.

Tableau 5.3 – Temps d'exécution (en secondes) d'une itération de l'échantillonnage de Gibbs marginalisé pour les modèles TAM, SN-LDA, VODUM, SNVDM-WII, SNVDM, SNVDM-GPU ($\tau = 10$) et SNVDM-GPU ($\tau = \infty$) sur la collection Indyref (avec $T = 10$) et Midterms (avec $T = 15$).

Modèle	Indyref	Midterms
TAM	1,45	0,87
SN-LDA	1,18	0,64
VODUM	2,78	1,85
SNVDM-WII	2,08	1,08
SNVDM	2,49	1,15
SNVDM-GPU ($\tau = 10$)	3,47	1,34
SNVDM-GPU ($\tau = \infty$)	14,67	2,56

5.3.2.3 Temps d'exécution

Nous discutons dans cette section de l'efficacité en terme de temps d'exécution des modèles proposés comparés aux modèles de référence. La machine sur laquelle ont été effectués les expérimentations est un ordinateur portable avec 8 CPU i7-4700MQ cadencés à 2,40 GHz (bien que notre implémentation ne soit pas parallélisée) et 8 Go de RAM. Nous rapportons dans le Tableau 5.3 le temps (en secondes) pris par une itération de l'échantillonnage de Gibbs marginalisé pour les différents modèles sur Indyref (avec $T = 10$) et Midterms (avec $T = 15$). Bien que nous observons une plus grande rapidité des implémentations de SN-LDA et TAM, les temps d'exécution de SNVDM et SNVDM-GPU ($\tau = 10$) ont le même ordre de magnitude : environ 2 et 3 fois plus lents au maximum, respectivement. Les temps d'exécutions sur Midterms sont très similaires pour tous les modèles, ce qui s'explique par le nombre d'interactions plus limités. Avec plus d'interactions, la différence est plus notable sur Indyref. En effet, les modèles basés sur SNVDM et en particulier ceux basés sur GPU ont un temps d'exécution fortement dépendant du nombre d'interactions. Cela est confirmée par le temps d'exécution très lent de SNVDM-GPU ($\tau = \infty$) sur Indyref. Au contraire, le modèle SNVDM-GPU ($\tau = 10$) tourne en un temps raisonnable dû à la sélection plus restrictive des accointances utilisées dans GPU. Par conséquent, SNVDM-GPU ($\tau = 10$) constitue un bon compris entre performance et temps d'exécution.

5.3.2.4 Analyse qualitative des thèmes et des points de vue découverts

Dans la littérature des modèles thématiques, il est d'usage d'évaluer la qualité des thèmes découverts par un modèle en analysant la cohérence des mots les plus probables des différentes distributions de mots associées aux thèmes. Comme nous l'avons mentionné dans la Section 3.4.2, des métriques de cohérence thématique ont alors été développées pour comparer de manière quantitative la qualité des thèmes découverts par différents modèles. Cependant, les modèles proposés et les modèles de référence que nous souhaitons comparer ici ne possèdent pas les mêmes distributions de mots. Par exemple, SN-LDA présente uniquement des

distributions de mots spécifiques aux thèmes, tandis que les modèles dérivant de SNVDM distinguent quatre types de mots, comme décrit dans la Section 5.1 : des mots de fond (tirés de ϕ_{00}), des mots de point de vue (tirés de ϕ_{01}), des mots thématiques (tirés de ϕ_{10}) et des mots de point de vue thématiques (tirés de ϕ_{11}). Par conséquent, les métriques de cohérence thématique ne peuvent être appliquées dans notre cas pour comparer exhaustivement la qualité des thèmes découverts par les modèles que nous proposons et les modèles de référence. Pour cette raison, nous choisissons de fournir dans cette section une analyse qualitative des thèmes et points de vue découverts par notre approche, afin de valider le fait que les variables latentes modélisées reflètent réellement les dimensions souhaitées.

Il est en particulier intéressant d'étudier pour un thème donné les mots thématiques ainsi que les mots de point de vue thématiques associés : cela permet la comparaison entre le lexique neutre et le lexique subjectif adoptés pour un même thème. Ainsi, nous sélectionnons un thème découvert par notre modèle le plus performant SNVDM-GPU ($\tau = \infty$) sur chaque collection, Indyref (avec $T = 10$) et Midterms (avec $T = 15$). Nous indiquons les 20 mots thématiques et les 20 mots de point de vue thématiques les plus probables associés pour un thème issu d'Indyref dans le Tableau 5.4 et de Midterms dans le Tableau 5.5. Notons que les étiquettes attribuées aux thèmes dans ces tableaux ont été choisies manuellement.

Le Tableau 5.4 montre les mots les plus probables associés à un thème central de la collection Indyref : la question de l'*indépendance de l'Écosse*. Comme anticipé, les mots thématiques découverts se focalisent essentiellement sur des aspects neutres tels que le référendum (*#indyref*, *vote*, *campaign*). Au contraire, les points de vue pro-Oui et pro-Non sont clairement reflétées par des mots spécifiques à ces positions. Les partisans du Oui utilisent des *hashtags* spécifiques tels que *#voteyes*, *#yes*, *#yesscot* et *#midlothiansaysyes*. Les partisans du Non utilisent quant à eux *#bettertogether* (slogan des pro-Non). De plus, les utilisateurs soutenant le Non semblent soulever le problème de la monnaie (*currency*) en cas de séparation (*separation*) entre l'Écosse et le Royaume Uni.

Dans le Tableau 5.5, nous avons rapporté les mots les plus probables au sujet du thème de l'*énergie et des ressources*, découvert dans la collection Midterms. Ce thème exhibe une différence frappante entre le discours démocrate et le discours républicain. Les démocrates parlent de problèmes environnementaux avec des *hashtags* tels que *#actonclimate*, *#climatechange*, *#earthday* et *#climate*. À l'inverse, les républicains se focalisent plutôt sur l'impact économique lié à ces problématiques, par exemple en matière d'emploi (*jobs*) et de factures à payer (*bills*). Globalement, nous observons que les différents thèmes et points de vue découverts – et les mots qui leur sont associés – sont raisonnablement cohérents, ce qui confirme (H4).

5.4 Discussions

Dans ce chapitre, nous avons introduit le modèle SNVDM, permettant de découvrir conjointement les points de vue des utilisateurs de réseaux sociaux et les thèmes abordés par ceux-ci. La méthode présentée s'appuie à la fois sur le contenu textuel généré par les

Tableau 5.4 – Listes des 20 mots thématiques et des 20 mots de point de vue thématiques les plus probables associés au thème manuellement étiqueté comme « indépendance de l'Écosse ». Les mots de point de vue thématiques sont donnés pour chaque point de vue : Oui ($v = 1$) et Non ($v = 2$). À droite de chaque mot, nous indiquons sa probabilité, fournie par les estimateurs $\hat{\boldsymbol{\phi}}_{10}$ et $\hat{\boldsymbol{\phi}}_{11}$ obtenus à l'issue de l'échantillonneur de Gibbs marginalisé.

Mots thématiques	$\hat{\phi}_{10zw}$	Mots de point de vue thématiques (Oui)	$\hat{\phi}_{11vzw}$	Mots de point de vue thématiques (Non)	$\hat{\phi}_{11vzw}$
#indyref	0,357	#voteyes	0,132	#indyref	0,100
scotland	0,068	yes	0,077	uk	0,061
independence	0,015	scotland	0,051	salmond	0,031
vote	0,015	independence	0,025	#bettertogether	0,027
campaign	0,014	westminster	0,019	#scotdecides	0,023
scottish	0,014	vote	0,016	separation	0,022
uk	0,011	independent	0,014	currency	0,018
people	0,009	country	0,012	thanks	0,016
future	0,009	#yes	0,012	today	0,016
independent	0,009	#scotland	0,012	say	0,015
oil	0,009	#indyscot	0,008	read	0,013
better	0,008	#yesscot	0,008	scots	0,012
business	0,008	better	0,008	#fmqs	0,011
new	0,007	#midlothiansaysyes	0,008	union	0,011
help	0,007	#scotdecides	0,008	new	0,011
share	0,006	people	0,008	saying	0,010
best	0,006	#bizforscotland	0,007	nationalists	0,010
make	0,006	powers	0,007	nhs	0,010
#yesscot	0,006	wealth	0,007	follow	0,009
says	0,005	nhs	0,007	separate	0,008

utilisateurs ainsi que sur les interactions sociales (entrantes et sortantes) existant entre les utilisateurs, que nous avons interprétées à la lumière du phénomène d'homophilie. Nous proposons également une extension de notre modèle, nommée SNVDM-GPU, basée sur les urnes de Pólya généralisées afin d'intégrer les accointances d'accointances des utilisateurs pour solutionner le problème de faible densité du réseau social (c'est-à-dire lorsque celui-ci comprend seulement un nombre limité d'interactions entre utilisateurs). Les modèles développés ont été évalués sur deux collections de *tweets* associés à des sujets politiques clivants. Les expérimentations menées ont montré la supériorité significative de nos modèles en terme de regroupement de points de vue, soulignant ainsi l'importance clé des interactions sociales pour la tâche d'identification des points de vue – une observation en accord avec les travaux antérieurs [Magdy *et al.*, 2016]. Le processus de GPU, favorisant l'exploitation des interactions, a permis à nos modèles d'être légèrement plus robustes dans le cas où le réseau est peu dense. En particulier, le modèle SNVDM-GPU ($\tau = 10$) présente un temps d'exécution comparable à celui des modèles de référence, constituant ainsi un bon compromis entre performance et temps d'exécution. Par ailleurs, les thèmes et les points de vue découverts se

Tableau 5.5 – Listes des 20 mots thématiques et des 20 mots de point de vue thématiques les plus probables associés au thème manuellement étiqueté comme « énergie et ressources ». Les mots de point de vue thématiques sont donnés pour chaque point de vue : démocrate (Dém., $v = 1$) et républicain (Rép., $v = 2$). À droite de chaque mot, nous indiquons sa probabilité, fournie par les estimateurs $\hat{\boldsymbol{\phi}}_{10}$ et $\hat{\boldsymbol{\phi}}_{11}$ obtenus à l'issue de l'échantillonneur de Gibbs marginalisé.

Mots thématiques	$\hat{\phi}_{10zw}$	Mots de point de vue thématiques (Dém.)	$\hat{\phi}_{11vzw}$	Mots de point de vue thématiques (Rép.)	$\hat{\phi}_{11vzw}$
energy	0,051	#actonclimate	0,037	#4jobs	0,061
house	0,043	climate	0,032	#obamacare	0,031
new	0,021	#p2	0,030	#jobs	0,029
gas	0,020	change	0,022	gop	0,028
natural	0,018	#climatechange	0,018	obama	0,018
#energy	0,014	clean	0,017	bills	0,017
#ff	0,011	oil	0,016	jobs	0,016
#kxl	0,011	energy	0,016	house	0,016
support	0,010	#gop	0,011	act	0,015
economic	0,009	seec	0,010	watch	0,014
#hcr	0,009	#earthday	0,010	#keystonexl	0,014
america	0,009	pollution	0,009	plan	0,013
report	0,008	#climate	0,009	#gop	0,012
rep	0,008	member	0,007	president	0,012
action	0,008	weather	0,007	american	0,011
protect	0,008	amendment	0,007	address	0,011
legislation	0,007	water	0,006	create	0,010
health	0,007	environment	0,006	potus	0,010
leaders	0,007	public	0,006	weekly	0,009
time	0,007	carbon	0,006	bipartisan	0,009

sont révélés comme étant raisonnablement cohérents, validant ainsi le fait que les variables latentes modélisées reflètent effectivement des points de vue et des thèmes.

Plusieurs directions s'offrent à nous pour étendre notre approche. L'une d'elle consisterait à intégrer la dimension temporelle au modèle dans la même veine que [Ren *et al.*, 2016]. Cela permettrait alors de pouvoir observer l'évolution des points de vue au cours du temps tout en constatant les thématiques émergeant au fur et à mesure du déroulement des événements (par exemple dans le cadre d'une élection présidentielle). Disposer de ce type de modélisations peut s'avérer précieux pour la construction automatique de résumés. Une autre dimension qu'il pourrait être également intéressant d'explorer est la position géographique des utilisateurs, comme cela a été fait dans [Hong *et al.*, 2012]. En effet, cette dernière peut aider dans certains cas à l'identification des points de vue. Par exemple, certaines villes américaines sont connues pour leur soutien au parti démocrate ou républicain ; un utilisateur situé dans la même ville ou dans le même quartier qu'un grand nombre de (disons) démocrates a alors possiblement plus de chance d'être influencés par ceux-ci.

Une autre piste que nous envisageons est la distinction de plusieurs types de thèmes. Les travaux actuels en fouille de point de vue (et les nôtres compris) supposent que chaque thème est discuté par l'ensemble des points de vue – ce qui est concrètement reflété dans SNVDM et SNVDM-GPU par la modélisation de distributions ϕ_{11vz} pour $z = 1, ..., T$. Or il semble envisageable que certains thèmes soient au contraire entièrement spécifiques à un point de vue. C'est ce que l'on constate dans le Tableau 5.5 si l'on considère la question du climat et de l'environnement comme un thème en soi. Par conséquent, il pourrait être judicieux de permettre à un modèle thématique intégrant les points de vue de faire la distinction entre les thèmes spécifiques aux points de vue et les thèmes « débattus », sur lesquels les différentes points de vue s'expriment. Cela pourrait également faciliter l'alignement et la comparaison du discours spécifique aux points de vue, lorsque cette comparaison a lieu d'être (c'est-à-dire lorsque les thèmes sont débattus).

Enfin, nous souhaitons proposer une approche basée sur nos modèles afin de permettre ou d'améliorer l'identification du point de vue des utilisateurs isolés ou avec peu d'interactions. En effet, une des limites de SNVDM et SNVDM-GPU est de demeurer relativement dépendant de la densité du réseau social et du nombre d'interactions entre utilisateurs, comme nous l'avons vu dans la Section 5.3.2.2. Une solution pour pouvoir identifier plus efficacement le point de vue des utilisateurs avec peu d'interactions serait de procéder par *bootstrapping*. La première étape consisterait à regrouper en fonction de leur point de vue (en utilisant nos modèles SNVDM ou SNVDM-GPU) les utilisateurs avec des connexions denses – pour lesquels on a une plus grande confiance d'identifier correctement le point de vue. Ensuite, un classifieur serait entraîné à partir du contenu textuel de ces utilisateurs connectés et de leurs étiquettes bruitées attribuées lors de la première étape, puis appliqué aux utilisateurs isolés – dont la seule information exploitable est le texte généré – pour prédire leur point de vue. Cette méthode hybride pourrait alors potentiellement améliorer l'identification du point de vue des utilisateurs avec peu d'interactions en conservant une nature non supervisée.

Conclusion générale

Ce chapitre clôture le mémoire en résumant les contributions apportées par cette thèse et en proposant différentes pistes à explorer pour étendre nos travaux. Le résumé de nos contributions est présenté dans la Section 1. Les perspectives et travaux futurs envisagés sont ensuite détaillés dans la Section 2.

1 Résumé des contributions

Les contributions présentées dans ce mémoire s'inscrivent dans le contexte de la fouille de points de vue, qui permet de généraliser les opinions positives et négatives à des opinions plus subtiles telles que celles révélant une inclinaison politique. Afin de limiter la dépendance vis-à-vis de l'annotation de données, nous nous sommes en particulier focalisés sur des méthodes non supervisées pour la fouille de points de vue en nous appuyant sur des modèles thématiques. Nous avons ainsi proposé deux modèles intégrant conjointement les thèmes et les points de vue pour découvrir les points de vue sur les Web.

Dans notre première proposition, nous nous sommes intéressé aux textes ne comprenant aucune métadonnée (par exemple, métadonnée sociale), tels que les essais. La seule information exploitable dans ce contexte est alors le texte lui-même. Par conséquent, nous avons étudié dans quelle mesure l'indication linguistique fournie par les parties de discours s'avère utile pour la découverte de points de vue. En particulier, nous avons utilisé les parties de discours pour différencier mots d'opinion – dépendant d'un point de vue et d'un thème – et mots thématiques – dépendant d'un thème uniquement – en suivant une pratique similaire en fouille d'opinions individuelles. La contribution principale apportée par notre premier modèle VODUM (*viewpoint and opinion discovery unification model*) a ainsi été d'analyser si l'exploitation des parties de discours est bénéfique en fouille de points de vue, comme elle l'est en fouille d'opinions individuelles. Nous avons également exploré l'impact d'autres propriétés du modèle VODUM : l'attribution de thèmes au niveau des phrases (plutôt qu'au niveau des mots) et de points de vue au niveau des documents, et la dépendance des distributions de thème vis-à-vis des points de vue (plutôt que vis-à-vis des documents).

Afin de tester les différentes hypothèses sous-jacentes à ces propriétés, nous avons évalué notre modèle sur la collection Bitterlemons, contenant des essais portant sur le conflit israélo-palestinien. Les expérimentations réalisées montrent que notre modèle se compare positivement à l'état de l'art – représenté par les modèles LDA (*latent Dirichlet allocation*), JTV (*joint topic viewpoint*) et TAM (*topic aspect model*) – en termes de regroupement des points de vue. La comparaison de VODUM avec des versions dégénérées de ce modèle – chacune occultant une des propriétés de VODUM – a par ailleurs permis d'observer les bénéfices des différentes propriétés de manière isolée. Nous avons observé que la séparation des mots d'opinion et mots thématiques ainsi que l'attribution de thèmes au niveau de la phrase sont les

propriétés les plus bénéfiques à VODUM dans le cas étudié. La dépendance des distributions de thèmes vis-à-vis des points de vue et l'attribution de points de vue au niveau du document ont également contribué à la performance de VODUM, quoiqu'avec un impact moindre. D'autre part, nous avons illustré par une étude qualitative que les thèmes et points de vue obtenus par VODUM semblent intuitivement cohérents.

Notre seconde contribution est, quant à elle, focalisée sur les points de vue exprimés par les utilisateurs de réseaux sociaux. Nous proposons de prendre en compte à la fois le texte généré par ceux-ci ainsi que leurs interactions sociales (telles que le *retweet* et la *mention* sur Twitter) pour découvrir leurs points de vue. L'utilisation des interactions est motivée par le principe d'homophilie, selon lequel les utilisateurs similaires (entre autres par leur point de vue) ont une plus grande propension à interagir. Nous décrivons tout d'abord le modèle SNVDM combinant deux modèles antérieurs : TAM et SN-LDA (*social network latent Dirichlet allocation*). Le premier permet la modélisation des points de vue et des thèmes à partir du texte tandis que le second exploite texte et interactions sociales pour la découverte de communautés. SNVDM présente par ailleurs l'originalité de prendre en compte les interactions sortantes et entrantes d'un utilisateur (respectivement, celles que l'utilisateur a initiées et celles dont il a été la cible). Dans la littérature, seules ces premières interactions sont utilisées, alors que ces dernières peuvent également être riches en information pour un utilisateur qui est connecté principalement ou uniquement par des interactions entrantes (par exemple dans le cas d'une personnalité qui est souvent *retweetée* du fait de sa popularité, mais qui ne *retweete* que rarement).

Nous avons ensuite développé une extension de SNVDM, intitulée SNVDM-GPU, basée sur les urnes de Pólya généralisées (GPU). Dans ce modèle, nous avons cherché à limiter l'impact de la faible densité des réseaux sociaux en exploitant les liens plus faibles existant entre un utilisateur et les « accointances de ses accointances ». Intuitivement, l'utilisation d'urnes de Pólya généralisées rend alors possible la prise en compte de ces interactions de second ordre lors de l'attribution d'un point de vue à un utilisateur. À notre connaissance, SNVDM-GPU est le premier modèle à appliquer le processus d'urnes de Pólya généralisées aux réseaux sociaux – les travaux précédents exploitant ce processus uniquement pour intégrer dans un modèle thématique la similarité sémantique entre les mots.

Les expérimentations conduites dans le cadre de notre seconde contribution ont comparé SNVDM et SNVDM-GPU aux modèles de l'état de l'art VODUM, TAM et SN-LDA sur deux collections issues de Twitter. Les résultats obtenus montrent la supériorité significative de nos modèles en terme de regroupement de points de vue, mettant ainsi en avant l'importance clé des interactions sociales pour l'identification des points de vue. De plus, nous observons que le processus de GPU, qui augmente l'impact des interactions, a permis à notre approche d'être légèrement plus robuste dans le cas où le réseau social est peu dense. Par ailleurs, l'étude du temps d'exécution des différents modèles a montré que le processus GPU peut être intégré sans entraîner un surcoût prohibitif en terme de temps. Enfin, la cohérence raisonnable des thèmes et des points de vue obtenus – analysés qualitativement – a permis de valider le fait que les variables latentes modélisées capturent les dimensions souhaitées.

Nous soulignons également que l'implémentation des différents modèles proposés dans nos contributions a été rendu publiquement accessible [7] [8] (sous forme de code source et d'exécutable en ligne de commande) afin de faciliter la reproduction de nos résultats et pour permettre leur comparaison à de futurs travaux. Nous souhaitons, par cet effort, favoriser le développement de nouvelles approches pour la fouille de points de vue.

2 Perspectives et travaux futurs

Les travaux rapportés dans ce mémoire pourraient être étendus selon plusieurs directions. Tout d'abord, il serait intéressant d'apporter des validations supplémentaires de nos modèles (ainsi que des autres modèles existants) dans le cas où le nombre de points de vue n'est pas limité à deux. Par exemple, dans le contexte de la politique, les points de vue pourraient refléter un soutien à un parti (par exemple, « La République En Marche », le « Front National », « La France Insoumise » et « Les Républicains ») plutôt qu'à un bord politique (gauche ou droite). De manière équivalente, cela reviendrait à sectionner l'axe politique gauche-droite en différents tronçons représentant chacun un parti. La tâche d'identification de points de vue nuancés a été essentiellement ignorée dans la littérature. Ce n'est que très récemment, dans [Preotiuc-Pietro *et al.*, 2017], que ce problème a été abordé. Toutefois, l'approche proposée demeure supervisée.

Deux solutions, basées sur des modèles non supervisés, sont néanmoins envisageables pour traiter ce problème : simplement augmenter le nombre de points de vue afin d'intégrer les différentes nuances de points de vue ou modéliser les points de vue par des variables latentes réelles pour capturer les corrélations entre les points de vue « proches » sur l'axe politique gauche-droite. Cette deuxième solution se rapproche de travaux plus anciens tels que la méthode NOMINATE [Poole et Rosenthal, 1985] et des modèles d'*ideal point* [Barberá, 2015; Gerrish et Blei, 2011, 2012; Nguyen *et al.*, 2015] – toutefois principalement appliqués aux points de vue des législateurs et donc peu généralisables aux utilisateurs « lambdas » des réseaux sociaux. Il pourrait alors être intéressant d'envisager des modèles plus généralement applicables intégrant à la fois texte et interactions sociales pour situer les points de vue des utilisateurs de réseaux sociaux avec plus de nuances.

Une autre direction de recherche que nous souhaitons explorer est l'intégration d'autres métadonnées, telles que l'horodatage du contenu générés par les utilisateurs de médias sociaux. De nombreux travaux ont étudié et modélisé la dynamique des thèmes – par exemple, [Blei et Lafferty, 2006; Wang et McCallum, 2006] – mais l'évolution des points de vue demeure un sujet non abordé dans la littérature, à notre connaissance. Les travaux de Ren *et al.* [2016] et de Liu *et al.* [2016, 2014] se sont intéresses à des problématiques proches quoique différentes. Bien qu'intégrant la temporalité des points de vue, Ren *et al.* [2016] se basent sur une notion de point de vue différente : les points de vue mentionnés sont en réalité des couples contenant chacun un thème et une polarité – se rapportant alors plutôt à ce que

7. `https://github.com/tthonet/VODUM`
8. `https://github.com/tthonet/SNVDM`

nous avons nommé « opinions individuelles » dans ce mémoire. Pour leur part, Liu *et al.* [2016, 2014] étudient la dynamique des communautés thématiques sur les réseaux sociaux. Les communautés thématiques sont clairement distinctes des communautés unies par un point de vue : les premières discutent d'un sujet qui leur est propre alors que les secondes débattent entre elles sur des thèmes communs. Étudier et modéliser l'évolution des points de vue, par exemple dans le cadre d'une élection présidentielle, demeure donc un problème ouvert.

Sur le long terme, nous envisageons d'utiliser les modèles que nous proposons pour construire des résumés de points de vue dénués du biais autrement inhérent à leurs auteurs. En effet, si un résumé de points de vue était constitué par un humain, ce dernier serait sans doute influencé par son propre point de vue lors de la rédaction ou de la collection des différentes parties du résumé. Il pourrait, consciemment ou non, être tenté de choisir des arguments ou opinions plus avantageux vis-à-vis de ses propres idées. Une forme possible du résumé de points de vue à constituer est la carte argumentative, qui organise les différents arguments ou opinions à la fois suivant les différents points de vue existants pour un problème donné – un sujet controversé – et suivant les thèmes sous-jacents à ce problème. De tels résumés nécessitent également d'être évalués par des utilisateurs réels (par exemple dans le cadre de la consommation d'articles de presse en ligne) afin de valider leur informativité et de comparer le biais des résumés générés automatiquement à ceux écrits par des humains.

Bibliographie

Abu-Jbara, A., Dasigi, P., Diab, M. et Radev, D. (2012). Subgroup Detection in Ideological Discussions. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL '12, pages 399–409.

Ahmed, A. et Xing, E. P. (2010). Staying Informed: Supervised and Semi-Supervised Multi-View Topical Analysis of Ideological Perspective. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 1140–1150.

Akoglu, L. (2014). Quantifying Political Polarity Based on Bipartite Opinion Networks. *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM '14, pages 2–11.

Al Zamal, F., Liu, W. et Ruths, D. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, ICWSM '12, pages 387–390.

Aldous, D. J. (1985). Exchangeability and Related Topics. *In École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198.

Allcott, H. et Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. Rapport technique, National Bureau of Economic Research.

Amigó, E., Gonzalo, J., Artiles, J. et Verdejo, F. (2009). A Comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints. *Information Retrieval*, 12(4): 461–486.

Asuncion, A., Welling, M., Smyth, P. et Teh, Y. W. (2009). On Smoothing and Inference for Topic Models. *In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34.

Augenstein, I., Rocktäschel, T., Vlachos, A. et Bontcheva, K. (2016). Stance Detection with Bidirectional Conditional Encoding. *In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 876–885.

Awadallah, R., Ramanath, M. et Weikum, G. (2012). PolariCQ: Polarity Classification of Political Quotations. *In Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1945–1949.

Baccianella, S., Esuli, A. et Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet. *In Proceedings of the International Conference on Language Resources and Evaluation*, LREC '10, pages 2200–2204.

Baeza-Yates, R. et Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. ACM Press / Addison-Wesley.

Bakshy, E., Messing, S. et Adamic, L. (2015). Exposure to Ideologically Diverse News and Opinion on Facebook. *Science*, 348(6239):1130–1132.

Balasubramanyan, R., Cohen, W. W., Pierce, D. et Redlawsk, D. P. (2012). Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News? *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media Modeling*, ICWSM '12, pages 18–25.

Bamman, D. et Smith, N. A. (2015). Open Extraction of Fine-Grained Political Statements. *In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 76–85.

Barberá, P. (2015). Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis*, 23(1):76–91.

Barbieri, N., Bonchi, F. et Manco, G. (2014). Who to Follow and Why: Link Prediction with Explanations. *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1266–1275.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M. et Lafferty, J. D. (2006). Dynamic topic models. *In Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120.

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2001). Latent Dirichlet Allocation. *In Proceedings of the 15th Annual Conference on Neural Information Processing Systems*, NIPS '01, pages 601–608.

Blei, D. M., Ng, A. Y. et Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boyd-Graber, J., Hu, Y. et Mimno, D. (2017). Applications of Topic Models. *Foundations and Trends in Information Retrieval*, 11(2):143–296.

Brigadir, I., Greene, D. et Cunningham, P. (2015). Analyzing Discourse Communities with Distributional Semantic Models. *In Proceedings of the 2015 ACM Conference on Web Science*, WebSci '15.

Brody, S. et Elhadad, N. (2010). An Unsupervised Aspect-Sentiment Model for Online Reviews. *In Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, NAACL-HLT '10, pages 804–812.

Cabrio, E. et Villata, S. (2012). Combining Textual Entailment and Argumentation Theory for Supporting Online Debates Interactions. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, ACL '12, pages 208–212.

Cardie, C., éditeur (2015). *Proceedings of the 2nd Workshop on Argumentation Mining*. ArgMining@NAACL-HLT '15. Association for Computational Linguistics.

Carenini, G., Ng, R. T. et Pauls, A. (2006). Interactive Multimedia Summaries of Evaluative Text. *In Proceedings of the 11th International Conference on Intelligent User Interfaces*, IUI '06, pages 124–131.

Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C. et Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, NIPS '09, pages 288–296.

Chen, C., Buntine, W., Ding, N., Xie, L. et Du, L. (2015). Differential Topic Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):230–242.

Chen, Z. et Liu, B. (2014). Mining Topics in Documents: Standing on the Shoulders of Big Data. *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 1116–1125.

Cohen, R. et Ruths, D. (2013). Classifying Political Orientation on Twitter: It's Not Easy! *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM '13, pages 91–99.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A. et Menczer, F. (2011a). Predicting the Political Alignment of Twitter Users. *In Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing*, PASSAT/SocialCom '11, pages 192–199.

Conover, M. D., Ratkiewicz, J., Francisco, M. R., Gonçalves, B., Menczer, F. et Flammini, A. (2011b). Political Polarization on Twitter. *In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 89–96.

Croutte, P. et Lautié, S. (2016). Le Baromètre du Numérique 2016. Rapport technique, Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie.

Dave, K., Lawrence, S. et Pennock, D. M. (2003). Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. *In Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. et Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Ding, X., Liu, B. et Yu, P. S. (2008). A Holistic Lexicon-Based Approach to Opinion Mining. *In Proceedings of the International Conference on Web Search and Data Mining*, WSDM '08, pages 231–239.

Dori-Hacohen, S. et Allan, J. (2013). Detecting Controversy on the Web. *In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 1845–1848.

Dori-Hacohen, S. et Allan, J. (2015). Automated Controversy Detection on the Web. *In Proceedings of the 37th European Conference on IR Research*, ECIR '15, pages 423–434.

Duggan, M. et Smith, A. (2016). The Political Environment on Social Media. Rapport technique, Pew Research Center.

Dunn, A. G., Leask, J., Zhou, X., Mandl, K. D. et Coiera, E. (2015). Associations Between Exposure to and Expression of Negative Opinions About Human Papillomavirus Vaccines on Social Media: An Observational Study. *Journal of Medical Internet Research*, 17(6):e144.

Escobar, M. D. et West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Fang, A., Ounis, I., Habel, P., Macdonald, C. et Limsopatham, N. (2015). Topic-centric Classification of Twitter User's Political Orientation. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 791–794.

Fang, L., Qian, Q., Huang, M. et Zhu, X. (2014). Ranking Sentiment Explanations for Review Summarization Using Dual Decomposition. *In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM '14, pages 1931–1934.

Fang, Y., Si, L., Somasundaram, N. et Yu, Z. (2012). Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model. *In Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, WSDM '12, pages 63–72.

Flaxman, S. R., Goel, S. et Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320.

Galam, S. (2017). Marine Le Pen Can Breach Her Glass Ceiling: The Drastic Effect of Differentiated Abstention.

Galley, M., McKeown, K., Hirschberg, J. et Shriberg, E. (2004). Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, pages 669–676.

Gamon, M. (2004). Sentiment Classification on Customer Feedback Data: Noisy Data, Large Feature Vectors, and the Role of Linguistic Analysis. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 841–847.

Gamon, M., Aue, A., Corston-Oliver, S. et Ringger, E. (2005). Pulse: Mining Customer Opinions from Free Text. *In Proceedings of the 2005 International Symposium on Intelligent Data Analysis*, IDA '05, pages 121–132.

Ganesan, K. et Zhai, C. (2012). Micropinion Generation: An Unsupervised Approach to Generating Ultra-Concise Summaries of Opinions. *In Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 869–878.

Ganesan, K., Zhai, C. et Han, J. (2010). Opinosis: A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions. *In Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 340–348.

Gao, H., Mahmud, J., Chen, J., Nichols, J. et Zhou, M. (2014). Modeling User Attitude toward Controversial Topics in Online Social Media. *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media*, ICWSM '14, pages 121–130.

Garimella, K., De Francisci Morales, G., Gionis, A. et Mathioudakis, M. (2016). Quantifying Controversy in Social Media. *In Proceedings of the 9th ACM International Conference on Web Search and Data Mining*, WSDM '16, pages 33–42.

Garimella, K., De Francisci Morales, G., Gionis, A. et Mathioudakis, M. (2017). Reducing Controversy by Connecting Opposing Views. *In Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 81–90.

Garrett, R. K. et Stroud, N. J. (2014). Partisan Paths to Exposure Diversity: Differences in Pro- and Counterattitudinal News Consumption. *Journal of Communication*, 64(4):680–701.

Gayo-Avello, D. (2011). Don't Turn Social Media into another 'Literary Digest' Poll. *Communications of the ACM*, 54(10):121–128.

Gayo-Avello, D. (2012). No, You Cannot Predict Elections with Twitter. *IEEE Internet Computing*, 16(6):91–94.

Gerrish, S. et Blei, D. (2011). Predicting Legislative Roll Calls from Text. *In Proceedings of the 28th International Conference on Machine Learning*, ICML '11, pages 489–496.

Gerrish, S. M. et Blei, D. M. (2012). How they Vote: Issue-Adjusted Models of Legislative Behavior. *In Proceedings of the 26th Annual Conference on Neural Information Processing Systems*, NIPS '12, pages 2762–2770.

Goldberg, A. et Zhu, X. (2006). Seeing Stars When There Aren't Many Stars: Graph-based Semi-supervised Learning for Sentiment Categorization. *In Proceedings of the 1st Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs@NAACL-HLT '06, pages 45–52.

Gottipati, S., Qiu, M., Sim, Y., Jiang, J. et Smith, N. A. (2013). Learning Topics and Positions from Debatepedia. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP '13, pages 1858–1868.

Gottipati, S., Qiu, M., Yang, L., Zhu, F. et Jiang, J. (2014). An Integrated Model For User Attribute Discovery: A Case Study on Political Affiliation Identification. *In Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD '14, pages 434–446.

Graells-Garrido, E., Lalmas, M. et Baeza-Yates, R. (2015). Finding Intermediary Topics Between People of Opposing Views: A Case Study.

Green, N., Ashley, K., Litman, D., Reed, C. et Walker, V., éditeurs (2014). *Proceedings of the 1st Workshop on Argumentation Mining*. ArgMining@ACL '14. Association for Computational Linguistics.

Griffiths, T. L. et Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl. 1):5228–5235.

Guerra, P. H. C., Meira Jr., W., Cardie, C. et Kleinberg, R. (2013). A Measure of Polarization on Social Media Networks Based on Community Boundaries. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, ICWSM '13, pages 215–224.

Guo, J., Lu, Y., Mori, T. et Blake, C. (2015). Expert-Guided Contrastive Opinion Summarization for Controversial Issues. *In Proceedings of the 24th International Conference Companion on World Wide Web*, WWW '15 Companion, pages 1105–1110.

Hardisty, E. A., Boyd-Graber, J. et Resnik, P. (2010). Modeling Perspective using Adaptor Grammars. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 284–292.

Hatzivassiloglou, V. et McKeown, K. R. (1997). Predicting the Semantic Orientation of Adjectives. *In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, ACL/EACL '97, pages 174–181.

Hatzivassiloglou, V. et Wiebe, J. M. (2000). Effects of Adjective Orientation and Gradability on Sentence Subjectivity. *In Proceedings of the 18th Conference on Computational Linguistics*, COLING '00, pages 299–305.

He, Y., Lin, C., Gao, W. et Wong, K.-F. (2012). Tracking Sentiment and Topic Dynamics from Social Media. *In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, ICWSM'12, pages 483–486.

He, Y., Lin, C., Gao, W. et Wong, K.-F. (2013). Dynamic Joint Sentiment-Topic Model. *ACM Transactions on Intelligent Systems and Technology*, 5(1):6:1–21.

Heinrich, G. (2008). Parameter Estimation for Text Analysis. Rapport technique, Fraunhofer Institute for Computer Graphics.

Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57.

Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1/2):177–196.

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J. et Tsioutsiouliklis, K. (2012). Discovering Geographical Topics in the Twitter Stream. *In Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 769–778.

Hu, M. et Liu, B. (2004). Mining and Summarizing Customer Reviews. *In Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177.

Iyyer, M., Enns, P., Boyd-Graber, J. et Resnik, P. (2014). Political Ideology Detection Using Recursive Neural Networks. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL '14, pages 1113–1122.

Jang, M. et Allan, J. (2016). Improving Automated Controversy Detection on the Web. *In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, SIGIR '16, pages 865–868.

Jang, M., Foley, J., Dori-Hacohen, S. et Allan, J. (2016). Probabilistic Approaches to Controversy Detection. *In Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM '16, pages 2069–2072.

Jin, Z., Cao, J., Zhang, Y. et Luo, J. (2016). News Verification by Exploiting Conflicting Social Viewpoints in Microblogs. *In Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI '16, pages 2972–2978.

Jo, Y. et Oh, A. H. (2011). Aspect and Sentiment Unification Model for Online Review Analysis. *In Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 815–824.

Johnson, K. et Goldwasser, D. (2016). "All I Know about Politics Is What I Read in Twitter": Weakly Supervised Models for Extracting Politicians' Stances From Twitter. *In Proceedings of the 26th International Conference on Computational Linguistics*, COLING '16, pages 2966–2977.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. et Saul, L. K. (1999). An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233.

Joshi, A., Bhattacharyya, P. et Carman, M. (2016). Political Issue Extraction Model: A Novel Hierarchical Topic Model That Uses Tweets By Political And Non-Political Authors. *In Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, WASSA@NAACL-HLT '16, pages 82–90.

Karoui, J., Benamara Zitoune, F., Moriceau, V., Aussenac-Gilles, N. et Belguith, L. H. (2015). Towards a Contextual Pragmatic Model to Detect Irony in Tweets. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL/IJCNLP '15, pages 644–650.

Kim, A., Murphy, J., Richards, A., Hansen, H., Powell, R. et Haney, C. (2014). Can Tweets Replace Polls? A U.S. Health-Care Reform Case Study. *In Social Media, Sociality, and Survey Research*, chapitre 3, pages 61–86. John Wiley & Sons, Inc.

Kim, H. D., Ganesan, K., Sondhi, P. et Zhai, C. (2011). Comprehensive Review of Opinion Summarization. Rapport technique, University of Illinois at Urbana-Champaign.

Kim, H. D. et Zhai, C. (2009). Generating Comparative Summaries of Contradictory Opinions in Text. *In Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, CIKM '09, pages 385–393.

Kim, P. (2006). The Forrester Wave: Brand Monitoring. Rapport technique, Forrester Wave.

Kim, S., Zhang, J., Chen, Z., Oh, A. et Liu, S. (2013). A Hierarchical Aspect-Sentiment Model for Online Reviews. *In Proceedings of the 27th AAAI Conference on Artificial Intelligence*, AAAI '13, pages 526–533.

Kim, S.-M. et Hovy, E. (2004). Determining the Sentiment of Opinions. *In Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, pages 1367–1373.

Kiritchenko, S. et Mohammad, S. M. (2016). SemEval-2016 Task 7: Determining Sentiment Intensity of English and Arabic Phrases. *In Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT '16, pages 42–51.

Lampos, V., Preotiuc-Pietro, D. et Cohn, T. (2013). A User-Centric Model of Voting Intention from Social Media. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 993–1003.

Lau, J. H., Newman, D. et Baldwin, T. (2014). Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. *In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '14, pages 530–539.

Leskovec, J., Lang, K. J. et Mahoney, M. W. (2010). Empirical Comparison of Algorithms for Network Community Detection. *In Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 631–640.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E. et Slonim, N. (2014). Context Dependent Claim Detection. *In Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 1489–1500.

Lewenberg, Y., Bachrach, Y., Bordeaux, L. et Kohli, P. (2016). Political Dimensionality Estimation Using a Probabilistic Graphical Model. *In Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, UAI '16, pages 447–456.

Li, C., Xing, J., Sun, A. et Ma, Z. (2016). Effective Document Labeling with Very Few Seed Words: A Topic Model Approach. *In Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM '16, pages 85–94.

Lim, K. W. et Buntine, W. (2014). Twitter Opinion Topic Model: Extracting Product Opinions from Tweets by Leveraging Hashtags and Sentiment Lexicon. *In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM '14, pages 1319–1328.

Lin, C. et He, Y. (2009). Joint Sentiment/Topic Model for Sentiment Analysis. *In Proceedings of the 18th ACM International Conference on Information and Knowledge Management*, CIKM '09, pages 375–384.

Lin, C., He, Y., Everson, R. et Rüger, S. (2012). Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145.

Lin, W.-H. et Hauptmann, A. (2006). Are These Documents Written from Different Perspectives? A Test of Different Perspectives Based On Statistical Distribution Divergence. *In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, COLING/ACL '06, pages 1057–1064.

Lin, W.-H., Wilson, T., Wiebe, J. et Hauptmann, A. (2006). Which Side are You on? Identifying Perspectives at the Document and Sentence Levels. *In Proceedings of the 10th Conference on Computational Natural Language Learning*, CoNLL '06, pages 109–116.

Lin, W.-H., Xing, E. et Hauptmann, A. (2008). A Joint Topic and Perspective Model for Ideological Discourse. *In Proceedings of 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, ECML/PKDD '08, pages 17–32.

Lippi, M. et Torroni, P. (2015). Context-Independent Claim Detection for Argument Mining. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence*, IJCAI '15, pages 185–191.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool.

Liu, B., Hu, M. et Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. *In Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 342–351.

Liu, J. S. (1994). The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *Journal of the American Statistical Association*, 89(427): 958–966.

Liu, Z., Zheng, Q., Wang, F. et Qian, B. (2016). Nonparametric Models for Characterizing the Topical Communities in Social Network. *Neurocomputing*, 216:439–450.

Liu, Z., Zheng, Q., Wang, F., Tian, Z. et Li, B. (2014). A Dynamic Nonparametric Model for Characterizing the Topical Communities in Social Streams. *In Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 379–387.

Lu, Y. et Zhai, C. (2008). Opinion Integration Through Semi-supervised Topic Modeling. *In Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 121–130.

Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A. et Baldwin, T. (2016). #ISISisNotIslam or #DeportAllMuslims? Predicting Unspoken Views. *In Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 95–106.

Mahmoud, H. M. (2008). *Pólya Urn Models*. Chapman & Hall/CRC.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

McDonald, R., Hannan, K., Neylon, T., Wells, M. et Reynar, J. (2007). Structured Models for Fine-to-Coarse Sentiment Analysis. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, page 432.

Mei, Q., Ling, X., Wondra, M., Su, H. et Zhai, C. (2007). Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. *In Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 171–180.

Meng, X., Wei, F., Liu, X., Zhou, M., Li, S. et Wang, H. (2012). Entity-Centric Topic-Oriented Opinion Summarization in Twitter. *In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 379–387.

Menini, S. et Tonelli, S. (2016). Agreement and Disagreement: Comparison of Points of View in the Political Domain. *In Proceedings of the 26th International Conference on Computational Linguistics*, COLING '16, pages 2461–2470.

Mihalcea, R., Banea, C. et Wiebe, J. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, ACL '07, pages 976–983.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M. et McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272.

Mishne, G. (2006). Multiple Ranking Strategies for Opinion Retrieval in Blogs. *In Proceedings of the 15th Text Retrieval Conference*, TREC '06.

Moghaddam, S. (2010). Opinion Digger: An Unsupervised Opinion Miner from Unstructured Product Reviews. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1825–1828.

Moghaddam, S. et Ester, M. (2011). ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews. *In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 665–674.

Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X. et Cherry, C. (2016). SemEval-2016 Task 6: Detecting Stance in Tweets. *In Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT '16, pages 31–41.

Mukherjee, A. et Liu, B. (2012). Mining Contentions from Discussions and Debates. *In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 841–849.

Mukherjee, A. et Liu, B. (2013). Discovering User Interactions in Ideological Discussions. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, ACL '13, pages 671–681.

Murphy, J., Link, M. W., Childs, J. H., Tesfaye, C. L., Dean, E., Stern, M., Pasek, J., Cohen, J., Callegaro, M. et Harwood, P. (2014). Social Media in Public Opinion Research: Report of the AAPOR Task Force on Emerging Technologies in Public Opinion Research. Rapport technique, American Association for Public Opinion Research.

Nakov, P., Ritter, A., Rosenthal, S., Stoyanov, V. et Sebastiani, F. (2016). SemEval-2016 Task 4: Sentiment Analysis in Twitter. *In Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT '16, pages 1–18.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A. et Wilson, T. (2013). SemEval-2013 Task 2: Sentiment Analysis in Twitter. *In Proceedings of the 7th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT '13, pages 312–320.

Nasukawa, T. N. et Yi, J. (2003). Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *In Proceedings of the 2nd International Conference on Knowledge Capture*, K-CAP '03, pages 70–77.

Nenkova, A. et McKeown, K. (2011). Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Newman, D., Asuncion, A., Smyth, P. et Welling, M. (2009). Distributed Algorithms for Topic Models. *Journal of Machine Learning Research*, 10:1801–1828.

Newman, D., Lau, J. H., Grieser, K. et Baldwin, T. (2010). Automatic Evaluation of Topic Coherence. *In Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, NAACL-HLT '10, pages 100–108.

Newman, M. E. J. (2006). Modularity and Community Structure in Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–8582.

Nguyen, V.-A., Boyd-Graber, J., Resnik, P. et Miler, K. (2015). Tea Party in the House: A Hierarchical Ideal Point Topic Model and Its Application to Republican Legislators in the 112th Congress. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing*, ACL/IJCNLP '15, pages 1438–1448.

Oard, D., Elsayed, T., Wang, J., Wu, Y., Zhang, P., Abels, E., Lin, J. et Soergel, D. (2006). TREC-2006 at Maryland: Blog , Enterprise, Legal and QA Tracks. *In Proceedings of the 15th Text Retrieval Conference*, TREC '06.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. et Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 122–129.

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G. et Soboroff, I. (2006). Overview of the TREC-2006 Blog Track. *In Proceedings of the 15th Text Retrieval Conference*, TREC '06.

Pang, B. et Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, pages 271–278.

Pang, B. et Lee, L. (2005). Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales. *In Proceedings of the 43rd Annual Meeting of the ACL*, ACL '05, pages 115–124.

Pang, B. et Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Pang, B., Lee, L. et Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning Techniques. *In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, pages 79–86.

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You.* The Penguin Press.

Paul, M. J. et Girju, R. (2009). Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, EMNLP '09, pages 1408–1417.

Paul, M. J. et Girju, R. (2010). A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics. *In Proceedings of the 24th AAAI Conference on Artificial Intelligence*, AAAI '10, pages 545–550.

Paul, M. J., Zhai, C. et Girju, R. (2010). Summarizing Contrastive Viewpoints in Opinionated Text. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 66–76.

Pennacchiotti, M. et Popescu, A.-M. (2011). Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. *In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 430–438.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., María Jiménez-Zafra, S. et Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *In Proceedings of the 10th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT '16, pages 19–30.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. et Manandhar, S. (2014). Semeval-2014 Task 4: Aspect Based Sentiment Analysis. *In Proceedings of the 8th International Workshop on Semantic Evaluation*, SemEval '14, pages 27–35.

Poole, K. T. et Rosenthal, H. (1985). A Spatial Model for Legislative Roll Call Analysis. *American Journal of Political Science*, 29(2):357–384.

Popescu, A.-M. et Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. *In Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, HLT/EMNLP '05, pages 339–346.

Popescu, A.-M. et Pennacchiotti, M. (2010). Detecting Controversial Events from Twitter. *In Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1873–1876.

Preotiuc-Pietro, D., Liu, Y., Hopkins, D. J. et Ungar, L. (2017). Beyond Binary Labels: Political Ideology Prediction of Twitter Users. *In Proceedings of the 55th Conference of the Association for Computational Linguistics*, ACL '17, pages 729–740.

Qiu, M. (2015). *Mining User Viewpoints in Online Discussions*. Thèse de doctorat, Singapore Management University.

Qiu, M. et Jiang, J. (2013). A Latent Variable Model for Viewpoint Discovery from Threaded Forum Posts. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, NAACL-HLT '13, pages 1031–1040.

Qiu, M., Sim, Y., Smith, N. A. et Jiang, J. (2015). Modeling User Arguments, Interactions, and Attributes for Stance Prediction in Online Debate Forums. *In Proceedings of the 2015 SIAM International Conference on Data Mining*, SDM '15, pages 855–863.

Qiu, M., Yang, L. et Jiang, J. (2013a). Mining User Relations from Online Discussions using Sentiment Analysis and Probabilistic Matrix Factorization. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, NAACL-HLT '13, pages 401–410.

Qiu, M., Yang, L. et Jiang, J. (2013b). Modeling Interaction Features for Debate Side Clustering. *In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 873–878.

Rahman, M. M. et Wang, H. (2016). Hidden Topic Sentiment Model. *In Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 155–165.

Rao, A. et Spasojevic, N. (2016). Actionable and Political Text Classification using Word Embeddings and LSTM. *In Proceedings of the 5th International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM@KDD '16.

Reed, C., éditeur (2016). *Proceedings of the 3rd Workshop on Argument Mining*. ArgMining@ACL '16. Association for Computational Linguistics.

Ren, Z. et de Rijke, M. (2015). Summarizing Contrastive Themes via Hierarchical Non-Parametric Processes. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 93–102.

Ren, Z., Inel, O., Aroyo, L. et de Rijke, M. (2016). Time-aware Multi-Viewpoint Summarization of Multilingual Social Text Streams. *In Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, CIKM '16, pages 387–396.

Riloff, E. et Wiebe, J. (2003). Learning Extraction Patterns for Subjective Expressions. *In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 105–112.

Röder, M., Both, A. et Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *In Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A. et Stoyanov, V. (2015). SemEval-2015 Task 10: Sentiment Analysis in Twitter. *In Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval@NAACL-HLT '12, pages 451–463.

Russell, S. J. et Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Pearson Education.

Sachan, M., Contractor, D., Faruquie, T. A. et Subramaniam, L. V. (2012). Using Content and Interactions for Discovering Communities in Social Networks. *In Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 331–340.

Sachan, M., Dubey, A., Srivastava, S., Xing, E. P. et Hovy, E. (2014). Spatial Compactness meets Topical Consistency: Jointly modeling Links and Content for Community Detection. *In Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 503–512.

Schofield, A. et Mimno, D. (2016). Comparing Apples to Apple: The Effects of Stemmers on Topic Models. *Transactions of the Association of Computational Linguistics*, 4(1):287–300.

Shafiei, M. M. et Muios, E. E. (2006). Latent Dirichlet Co-Clustering. *In Proceedings of the 2006 IEEE International Conference on Data Mining*, ICDM '06, pages 542–551.

Smith, A. (2014). Cell Phones, Social Media and Campaign 2014. Rapport technique, Pew Research Center.

Smith, A. et Anderson, M. (2016). Online Shopping and E-Commerce. Rapport technique, Pew Research Center.

Snyder, B. et Barzilay, R. (2007). Multiple Aspect Ranking Using the Good Grief Algorithm. *In Proceedings of the 2007 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology*, NAACL-HLT '07, pages 300–307.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y. et Manning, C. D. (2011). Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 151–161.

Somasundaran, S. et Wiebe, J. (2010). Recognizing Stances in Ideological On-Line Debates. *In Proceedings of the 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, WCAAGET@NAACL-HLT '10, pages 116–124.

Strapparava, C. et Mihalcea, R. (2007). Semeval-2007 Task 14: Affective Text. *In Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval@ACL '07, pages 70–74.

Sunstein, C. R. (2009). *Republic.com 2.0*. Princeton University Press.

Tang, J., Yao, L. et Chen, D. (2009). Multi-topic based Query-oriented Summarization. *In Proceedings of the 2009 SIAM International Conference on Data Mining*, SDM '09, pages 1148–1159.

Teh, Y. W., Newman, D. et Welling, M. (2006). A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. *In Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, NIPS '06, pages 1353–1360.

Thonet, T., Cabanac, G., Boughanem, M. et Pinel-Sauvagnat, K. (2016). VODUM: A Topic Model Unifying Viewpoint, Topic and Opinion Discovery. *In Proceedings of the 38th European Conference on IR Research*, ECIR '16, pages 533–545.

Thonet, T., Cabanac, G., Boughanem, M. et Pinel-Sauvagnat, K. (2017). Users Are Known by the Company They Keep: Topic Models for Viewpoint Discovery in Social Networks. *In Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, CIKM '17.

Trabelsi, A. et Zaïane, O. R. (2014). Mining Contentious Documents Using an Unsupervised Topic Model Based Approach. *In Proceedings of the 2014 IEEE International Conference on Data Mining*, ICDM '14, pages 550–559.

Trabelsi, A. et Zaïane, O. R. (2015). Extraction and Clustering of Arguing Expressions in Contentious Text. *Data & Knowledge Engineering*, 100:226–239.

Trabelsi, A. et Zaïane, O. R. (2016). Mining Contentious Documents. *Knowledge and Information Systems*, 48(3):537–560.

Tsakalidis, A., Papadopoulos, S., Cristea, A. I. et Kompatsiaris, Y. (2015). Predicting Elections for Multiple Countries Using Twitter and Polls. *IEEE Intelligent Systems*, 30(2): 10–17.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. et Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, ICWSM '10, pages 178–185.

Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, ACL '02, pages 417–424.

Twardy, C. R. (2004). Argument Maps Improve Critical Thinking. *Teaching Philosophy*, 27(2):95–116.

van der Zwaan, J. M., Marx, M. et Kamps, J. (2016). Validating Cross-Perspective Topic Modeling for Extracting Political Parties' Positions from Parliamentary Proceedings. *In Proceedings of the 22nd European Conference on Artificial Intelligence*, ECAI '16, pages 28–36.

Wallach, H. M., Mimno, D. et McCallum, A. (2009). Rethinking LDA: Why Priors Matter. *In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, NIPS '09, pages 1973–1981.

Wang, L., Raghavan, H., Cardie, C. et Castelli, V. (2014). Query-Focused Opinion Summarization for User-Generated Content. *In Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, pages 1660–1669.

Wang, S., Chen, Z. et Liu, B. (2016). Mining Aspect-Specific Opinion using a Holistic Lifelong Topic Model. *In Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 167–176.

Wang, X. et McCallum, A. (2006). Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends. *In Proceedings of the 12nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433.

Wei, X. et Croft, W. B. (2006). LDA-based Document Models for Ad-Hoc Retrieval. *In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 178–185.

Wiebe, J. et Wilson, T. (2002). Learning to Disambiguate Potentially Subjective Expressions. *In Proceedings of the 6th Conference on Natural Language Learning*, CoNLL '02, pages 1–7.

Wiebe, J. M., Bruce, R. F. et O'Hara, T. P. (1999). Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. *In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, ACL '99, pages 246–253.

Wilson, T., Wiebe, J. et Hoffmann, P. (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *In Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT/EMNLP '05, pages 347–354.

Wu, Y. (2010). SemEval-2010 Task 18: Disambiguating Sentiment Ambiguous Adjectives. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval@ACL '10, pages 1191–1199.

Yang, K., Yu, N., Valerio, A. et Zhang, H. (2006). WIDIT in TREC-2006 Blog Track. *In Proceedings of the 15th Text Retrieval Conference*, TREC '06.

Yao, J.-g., Wan, X. et Xiao, J. (2017). Recent Advances in Document Summarization. *Knowledge and Information Systems*, 53(2):297–336.

Yu, H. et Hatzivassiloglou, V. (2003). Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. *In Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 129–136.

Zhang, W. et Yu, C. (2006). UIC at TREC 2006 Blog Track. *In Proceedings of the 15th Text Retrieval Conference*, TREC '06.

Zhao, W. X., Jiang, J., Yan, H. et Li, X. (2010). Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65.

Zhou, X., Coiera, E., Tsafnat, G., Arachi, D., Ong, M.-S. et Dunn, A. G. (2015). Using Social Connection Information to Improve Opinion Mining: Identifying Negative Sentiment about HPV Vaccines on Twitter. *In Proceedings of the 15th World Congress on Health and Biomedical Informatics*, MedInfo '15, pages 761–765.

Résumé — Les plateformes en ligne telles que les blogs et les réseaux sociaux permettent aux internautes de s'exprimer sur des sujets d'une grande variété (produits commerciaux, politique, services, etc.). Cet important volume de données d'opinions peut être exploré et exploité grâce à des techniques de fouille de texte connues sous le nom de fouille d'opinions ou analyse de sentiments. Contrairement à la majorité des travaux actuels en fouille d'opinions, qui se focalisent sur les opinions simplement positives ou négatives (ou un intermédiaire entre ces deux extrêmes), nous nous intéressons dans cette thèse aux points de vue. La fouille de point de vue généralise l'opinion au delà de son acception usuelle liée à la polarité (positive ou négative) et permet l'étude d'opinions exprimées plus subtilement, telles que les opinions politiques. Nous proposons dans cette thèse des approches non supervisées – ne nécessitant aucune annotation préalable – basées sur des modèles thématiques probabilistes afin de découvrir simultanément les thèmes et les points de vue exprimés dans des corpus de textes d'opinion. Dans notre première contribution, nous avons exploré l'idée de différencier mots d'opinions (spécifiques à la fois à un point de vue et à un thème) et mots thématiques (dépendants du thème mais neutres vis-à-vis des différents points de vue) en nous basant sur les parties de discours, inspirée par des pratiques similaires dans la littérature de fouille d'opinions classique – restreinte aux opinions positives et négatives. Notre seconde contribution se focalise quant à elle sur les points de vue exprimés sur les réseaux sociaux. Notre objectif est ici d'analyser dans quelle mesure l'utilisation des interactions entre utilisateurs, en outre de leur contenu textuel généré, est bénéfique à l'identification de leurs points de vue. Nos différentes contributions ont été évaluées et comparées à l'état de l'art sur des collections de documents réels.

Mots clés : fouille d'opinions, fouille de points de vue, modèles thématiques

Abstract — The advent of online platforms such as weblogs and social networking sites provided Internet users with an unprecedented means to express their opinions on a wide range of topics, including policy and commercial products. This large volume of opinionated data can be explored and exploited through text mining techniques known as opinion mining or sentiment analysis. Contrarily to traditional opinion mining work which mostly focuses on positive and negative opinions (or an intermediate in-between), we study a more challenging type of opinions: viewpoints. Viewpoint mining reaches beyond polarity-based opinions (positive/negative) and enables the analysis of more subtle opinions such as political opinions. In this thesis, we proposed unsupervised approaches – i.e., approaches which do not require any labeled data – based on probabilistic topic models to jointly discover topics and viewpoints expressed in opinionated data. In our first contribution, we explored the idea of separating opinion words (specific to both viewpoints and topics) from topical, neutral words based on parts of speech, inspired by similar practices in the litterature of non viewpoint-related opinion mining. Our second contribution tackles viewpoints expressed by social network users. We aimed to study to what extent social interactions between users – in addition to text content – can be beneficial to identify users' viewpoints. Our different contributions were evaluated and benchmarked against state-of-the-art baselines on real-world datasets.

Keywords: opinion mining, viewpoint mining, topic models

Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse CEDEX 9