Université Lumière Lyon 2
Ecole Doctorale Informatique et Mathématiques (ED 512)
Laboratoire ERIC (EA 3083)

THÈSE

Présentée par :
Mohamed Dermouche

Soutenue le 08/06/2015 pour obtenir le grade de :
Docteur de l'Université Lumière Lyon 2
Discipline : Informatique

Modélisation conjointe des thématiques et des opinions

Application à l'analyse des données textuelles issues du Web

Membres du jury :

Patrice Bellot	Professeur des universités, Aix-Marseille Université	Rapporteur
Osmar Zaïane	Professeur des universités, Université d'Alberta, Canada	Rapporteur
Patrick Gallinari	Professeur des universités, Université Pierre et Marie Curie, Paris	Examinateur
Mathieu Roche	Chercheur (HDR), CIRAD, Montpellier	Examinateur
Sabine Loudcher	Maître de conférences (HDR), Université Lumière Lyon 2	Directeur
Julien Velcin	Maître de conférences, Université Lumière Lyon 2	Co-directeur
Leila Khouas	Chercheur (docteur) AMI Software R&D, Montpellier	Invité

Abstract

This work is located at the junction of two domains : topic modeling and sentiment analysis. The problem that we propose to tackle is the joint and dynamic modeling of topics (subjects) and sentiments (opinions) on the Web. In the literature, the task is usually divided into sub-tasks that are treated separately. The models that operate this way fail to capture the topic-sentiment interaction and association. In this work, we propose a joint modeling of topics and sentiments, by taking into account associations between them. We are also interested in the dynamics of topic-sentiment associations.

To this end, we adopt a statistical approach based on the probabilistic topic models. Our main contributions can be summarized in two points :

1. TS (Topic-Sentiment model) : a new probabilistic topic model for the joint extraction of topics and sentiments. This model allows to characterize the extracted topics with distributions over the sentiment polarities. The goal is to discover the sentiment proportions specific to each of the extracted topics.
2. TTS (Time-aware Topic-Sentiment model) : a new probabilistic model to caracterize the topic-sentiment dynamics. Relying on the document's time information, TTS allows to characterize the quantitative evolution for each of the extracted topic-sentiment pairs.

We also present two other contributions : a new evaluation framework for measuring the performance of topic-extraction methods, and a new hybrid method for sentiment detection and classification from text. This method is based on combining supervised machine learning and prior knowledge. All of the proposed methods are tested on real-world data based on adapted evaluation frameworks.

Résumé

Cette thèse se situe à la confluence des domaines de "la modélisation de thématiques" (*topic modeling*) et "l'analyse d'opinions" (*opinion mining*). Le problème que nous traitons est la modélisation conjointe et dynamique des thématiques (sujets) et des opinions (prises de position) sur le Web et les médias sociaux. En effet, dans la littérature, ce problème est souvent décomposé en sous-tâches qui sont menées séparément. Ceci ne permet pas de prendre en compte les associations et les interactions entre les opinions et les thématiques sur lesquelles portent ces opinions (cibles). Dans cette thèse, nous nous intéressons à la modélisation conjointe et dynamique qui permet d'intégrer trois dimensions du texte (thématiques, opinions et temps).

Afin d'y parvenir, nous adoptons une approche statistique, plus précisément, une approche basée sur les modèles de thématiques probabilistes (*topic models*). Nos principales contributions peuvent être résumées en deux points :

1. Le modèle TS (*Topic-Sentiment model*) : un nouveau modèle probabiliste qui permet une modélisation conjointe des thématiques et des opinions. Ce modèle permet de caractériser les distributions d'opinion relativement aux thématiques. L'objectif est d'estimer, à partir d'une collection de documents, dans quelles proportions d'opinion les thématiques sont traitées.
2. Le modèle TTS (*Time-aware Topic-Sentiment model*) : un nouveau modèle probabiliste pour caractériser l'évolution temporelle des thématiques et des opinions. En s'appuyant sur l'information temporelle (date de création de documents), le modèle TTS permet de caractériser l'évolution des thématiques et des opinions quantitativement, c'est-à-dire en terme de la variation du volume de données à travers le temps.

Par ailleurs, nous apportons deux autres contributions : une nouvelle mesure pour évaluer et comparer les méthodes d'extraction de thématiques, ainsi qu'une nouvelle méthode hybride pour le classement d'opinions basée sur une combinaison de l'apprentissage automatique supervisé et la connaissance *a priori*. Toutes les méthodes proposées sont testées sur des données réelles en utilisant des évaluations adaptées.

Remerciements

Ce travail est dédié à mes parents.

Ce mémoire est le résultat d'un travail de plus de trois ans. En préambule, je veux adresser tous mes remerciements aux personnes avec lesquelles j'ai pu échanger et qui m'ont aidé de près ou de loin à bien mener ce travail.

En commençant par remercier tout d'abord mes trois directeurs de thèse **Mme. Leila Khouas**, **Mme. Sabine Loudcher** et **M. Julien Velcin**, sans qui cette thèse ne serait pas la moitié de ce qu'elle est. Merci pour l'encouragement et le suivi au quotidien.

Merci à **M. Eric Fourboul** de m'avoir confié ce passionnant sujet de thèse. Je remercie aussi l'ensemble de mes collègues montpelliérains d'AMI Software et tous mes collègues lyonnais du laboratoire ERIC.

Je remercie les membres du jury d'avoir accepté d'évaluer ce travail.

Merci à toute ma famille qui m'a toujours soutenu.

Notations

Les notations utilisées dans ce manuscrit sont données dans le tableau suivant :

Notation	Signification
$\mathbb{D}$	Un corpus (ensemble de documents)
D	Taille du corpus ($D = \|\mathbb{D}\|$)
d	Un document
n_d	Taille du document d (nombre de termes)
$\mathbb{V}$	Un vocabulaire (ensemble de termes)
V	Taille du vocabulaire ($V = \|\mathbb{V}\|$)
$\mathbb{Z}$	Un ensemble de thématiques
T	Nombre de thématiques ($T = \|\mathbb{Z}\|$)
z	Une thématique
w	Un terme (élément de $\mathbb{V}$)
c_+, c_-, c_0	Classes d'opinions : positive, négative et neutre respectivement
X	Matrice d'occurrences documents-termes représentant le corpus D
H	Matrice caractérisant les thématiques (espace latent)
W	Matrice de projection des documents dans l'espace latent
Φ	Matrice des distributions des thématiques sur le vocabulaire (φ)
Θ	Matrice des distributions des documents sur les thématiques (θ)
Π	Matrice des distributions des thématiques sur les polarités d'opinion (π)
Ψ	Matrice des distributions des paires thématiques-opinions sur les étiquettes temporelles (ψ)

TABLE 1 – Notations.

Sommaire

Chapitre 1

Introduction Générale

1.1 Contexte et problématiques

L'arrivée du Web 2.0 et son rapide essor grâce aux nouveaux modes de communication, comme les réseaux sociaux, les blogs et les forums de discussion, ont révolutionné nos habitudes de produire et de consommer l'information. En effet, le changement de la nature même du Web en un endroit où l'internaute est impliqué en tant que consommateur et producteur de l'information a conduit le Web à évoluer de manière très rapide. Prenons l'exemple du réseau social Twitter [1] où les internautes s'expriment sur la vie quotidienne en générant des quantités impressionnantes de données : environ 500 millions de tweets sont générés chaque jour à travers le monde [2]. Des millions d'utilisateurs diffusent leurs idées et leurs opinions sur des sujets divers et variés, comme les produits de consommation, les services, les hommes politiques, les événements de l'actualité, etc.

Dans ce contexte de grands volumes de données où l'information utile est encore enfouie et difficile à extraire, une exploration rapide et efficace de ces données doit reposer sur des outils de fouille de données et plus précisément de fouille de textes puisqu'on a majoritairement affaire à des données textuelles. Ces outils permettraient de mieux répondre à des besoins plus spécifiques des veilleurs (décideurs, spécialistes de communication et de marketing, etc.) tout en intégrant des contraintes d'analyse en terme de temps d'exécution et de volume de données. En effet, aujourd'hui, les veilleurs s'attendent à des outils puissants, capables de fournir des réponses à des questions telles que : quel aspect du nouveau produit ne plaît pas aux clients ? comment l'opinion publique par rapport à l'énergie nucléaire a-t-elle changé après l'incident nucléaire de Fukushima en 2011 ? et dans quelles proportions ?.. voire même à des questions de nature prédictif comme : quel serait l'impact d'une décision politique sur

1. http ://www.twitter.com
2. https ://blog.twitter.com/2013/new-tweets-per-second-record-and-how

l'opinion publique ?

La modélisation de thématiques et l'analyse d'opinions sont deux tâches de la fouille de textes qui permettent d'obtenir des éléments de réponse à ce type de questions. La modélisation de thématiques permet d'extraire les sujets saillants (de quoi parle-t-on ?) à partir d'une collection de documents textuels, alors que l'analyse d'opinions se focalise sur l'identification et la catégorisation des opinions et des prises de positions exprimées en langage naturel. Dans les dix dernières années, ces deux tâches de fouille de textes ont été intensivement étudiées par des chercheurs de domaines différents (informatique, linguistique, psychologie, sociologie).

Dans la littérature, le problème de modélisation conjointe et dynamique des thématiques et des opinions est souvent décomposé en sous-problèmes. Les travaux qui se concentrent sur la modélisation de thématiques ne prennent pas en compte l'association entre celles-ci et les opinions [9, 18, 44, 57]. Les travaux qui s'intéressent à l'analyse d'opinions ne font souvent pas assez de rapprochement avec les thématiques [25, 29, 51, 84, 85, 86]. Dans les cinq dernières années, un nombre important de travaux ont été proposés pour l'analyse conjointe des thématiques et des opinions soit en se basant sur une analyse linguistique fine pour la détection des cibles d'opinion, comme dans [45, 48, 65, 89, 98, 104, 106, 107, 125] ou en utilisant une modélisation probabiliste conjointe comme dans [37, 49, 52, 59, 60, 63, 64, 70, 115]. Cependant, ces travaux présentent deux inconvénients majeurs :

- aucun de ces travaux ne permet d'extraire les proportions d'opinions relatives aux thématiques, c'est-à-dire caractériser les thématiques en les associant à des opinions. Une telle connaissance serait très utile pour avoir une vue globale des relations thématiques-opinions.
- aucun de ces travaux ne permet de modéliser le changement de ces opinions à travers le temps. En effet, une thématique qui a été évoquée de manière positive dans une période de temps peut changer de polarité dans une autre période donnée.

Le problème auquel nous proposons de répondre dans cette thèse est la modélisation conjointe et dynamique des thématiques et des opinions sur le Web. L'objectif principal est de développer des modèles qui répondent à ces deux exigences tout en prenant en compte les interactions entre les thématiques et les opinions (modélisation conjointe) ainsi que l'influence de l'information temporelle (modélisation dynamique). Afin d'y parvenir, nous adoptons une approche probabiliste à base de modèles de thématiques (*topic models*). S'inscrivant principalement dans un contexte de veille sur le Web, nos travaux trouvent de nombreuses applications comme l'analyse de tendances, la e-réputation, la gestion de la relation avec les clients, et de manière plus générale l'analyse exploratoire des données textuelles sur les trois axes : thématiques,

opinions et temps.

Cette thèse s'est déroulée dans un contexte industriel (CIFRE [3]) au sein de l'entreprise AMI Software [4]. Celle-ci est spécialisée dans l'édition de logiciels de veille économique et stratégique, notamment la plateforme de veille sur le Web AMIEI (*AMI Enterprise Intelligence*). Les travaux réalisés dans cette thèse ont été intégrés au sein de la plateforme AMIEI afin de fournir des outils d'analyse axés sur la modélisation des thématiques et des opinions. Par ailleurs, ces travaux ont été réalisés sous la direction du laboratoire ERIC [5] de l'université Lumière Lyon 2, spécialisé dans la fouille de données et l'aide à la décision.

1.2 Contributions

La modélisation conjointe des thématiques et des opinions est un problème qui n'a été traité que partiellement dans la littérature. En effet, aucun des travaux existants ne permet d'estimer globalement les proportions des opinions relatives aux thématiques. De plus, aucun de ces travaux ne permet une modélisation dynamique qui prend en compte l'évolution temporelle des thématiques et des opinions simultanément. Dans cette thèse, nous proposons deux nouveaux modèles probabilistes afin de répondre à cette double problématique. Afin d'y parvenir, nous avons commencé par nous intéresser aux domaines de la modélisation de thématiques et celui de l'analyse d'opinions séparément. Cela nous a permis d'apporter deux autres contributions : une nouvelle mesure pour évaluer et comparer les méthodes d'extraction de thématiques et une nouvelle méthode hybride pour le classement d'opinions. L'ensemble de nos travaux peut être résumé en quatre contributions principales :

1. Une nouvelle mesure d'évaluation pour comparer différentes méthodes d'extraction de thématiques : pour caractériser les thématiques, ces méthodes utilisent différents paradigmes, comme des distributions de probabilités, des matrices ou des partitions de documents. L'hétérogénéité de ces résultats rend impossible leur comparaison dans le même cadre et de manière uniforme. Notre proposition pour résoudre ce problème consiste à projeter ces résultats dans un espace unifié qui permet de calculer une mesure d'évaluation et comparer ces résultats de manière quantitative et uniforme. Ces travaux ont été publiés dans [23].
2. Une nouvelle méthode hybride pour l'analyse d'opinions : dans cette contribution, nous nous intéressons au problème de sur-apprentissage

3. Conventions Industrielles de Formation par la REcherche
4. http ://www.amisw.com/
5. http ://eric.ish-lyon.cnrs.fr/

rencontré avec le classeur Bayésien naïf. En effet, cette méthode s'avère particulièrement sensible à la qualité de données d'apprentissage notamment quand ces données sont déséquilibrées ou insuffisamment représentatives. Afin d'atténuer ce problème, nous proposons une nouvelle méthode qui combine l'apprentissage automatique et la connaissance *a priori* exprimée sous forme d'un lexique d'opinions. Ces travaux ont été publiés dans [20].

3. TS (*Topic-Sentiment model*) : un nouveau modèle probabiliste qui permet une modélisation conjointe des thématiques et des opinions. Ce modèle permet de caractériser les proportions d'opinions relatives aux thématiques. L'objectif est d'estimer, à partir d'une collection de documents non annotés (apprentissage non supervisé) dans quelles proportions d'opinions les thématiques sont traitées. Ces travaux ont été publiés dans [21].
4. TTS (*Time-aware Topic-Sentiment model*) : un nouveau modèle probabiliste pour caractériser l'évolution temporelle des thématiques et des opinions. En s'appuyant sur l'information temporelle (date de création de documents), le modèle TTS permet de caractériser l'évolution des thématiques et des opinions quantitativement, c'est-à-dire en terme de volume de données, dans le temps. Ces travaux ont été publiés dans [22].

Tous ces travaux ont été testés et validés sur des jeux de données réelles en utilisant des évaluations adaptées. Ils ont été, par ailleurs, implémentés et intégrés au sein de la plateforme de veille AMIEI sous forme de composants d'analyse avancée. Une démonstration du composant de l'analyse d'opinions avec la méthode hybride a été publiée dans [19].

1.3 Organisation du manuscrit

Outre ce chapitre introductif, le manuscrit est organisé en six chapitres. Les chapitres 2 et 3 présentent nos travaux dans les domaines de "la modélisation de thématiques" et "l'analyse d'opinions" respectivement. Dans le chapitre 4, nous présentons nos travaux sur la modélisation conjointe des thématiques et des opinions (modèle TS). Dans le chapitre 5, nous présentons nos travaux sur l'évolution temporelle des thématiques et des opinions (modèle TTS).

Enfin, le chapitre 6 est consacré à la description du cadre industriel de la thèse ainsi que l'intégration de nos travaux au sein de la plateforme de veille AMIEI. Dans ce chapitre, nous aurons aussi l'occasion de démontrer la pertinence de la modélisation conjointe et dynamique à travers des cas d'étude sur des données réelles issues de l'actualité. Enfin, dans le chapitre 7, nous donnons une conclusion et quelques perspectives d'évolution de nos travaux.

Chapitre 2

Modélisation de Thématiques

Résumé. *Dans ce chapitre, nous nous intéressons à la tâche de modélisation de thématiques. Nous exposons les différentes problématiques liée à cette tâche (extraction, nommage, évaluation) et les méthodes associées. Nous soulignons les problèmes rencontrés pour évaluer les méthodes d'extraction de thématiques, en particulier l'absence d'une mesure qui permet l'évaluation et la comparaison de ces méthodes. Nous proposons une première solution à ce problème qui consiste en une nouvelle mesure de qualité, intitulée "Vraisemblance Généralisée", qui permet d'évaluer et de comparer différentes méthodes d'extraction de thématiques. Cette mesure est inspirée de la mesure de vraisemblance qui est largement utilisée dans le domaine de modélisation probabiliste.*

Sommaire

2.1 Introduction

Depuis plusieurs années, les données textuelles collectées sur le Web suscitent un intérêt croissant pour l'analyse exploratoire et/ou prédictive (e.g., recommandation automatique, analyse d'opinions, modélisation de thématiques, etc.). La modélisation de thématiques est une tâche qui consiste à extraire et caractériser les sujets, parfois appelés "structures thématiques latentes" à partir d'une collection de documents.

Le problème consiste, à partir d'un ensemble de données textuelles, à extraire les sujets saillants, que nous appelons ici *thématiques*. A notre connaissance, il n'y a pas à ce jour un consensus pour définir une thématique de manière formelle. La difficulté de donner une telle définition est accentuée par l'hétérogénéité des méthodes utilisées dans cette finalité. En effet, une thématique peut prendre plusieurs formes dont une distribution de probabilités, un graphe ou encore une liste de mots [9, 18, 32, 44, 57]. Dans notre travail, nous proposons la définition suivante :

Définition 1 (thématique) *Une thématique est un sujet unique et clairement identifiable dans un ou plusieurs documents.*

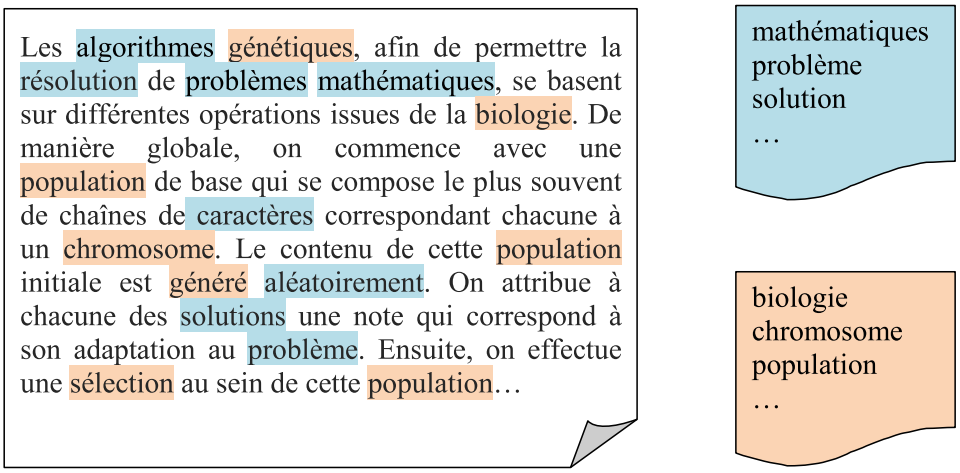


FIGURE 2.1.1 – Illustration de thématiques sur un texte décrivant les algorithmes génétiques.

La figure 2.1.1 représente un extrait de texte relatif aux algorithmes génétiques. Dans ce texte, nous distinguons deux thématiques majeures, celle des mathématiques et celle de la biologie. Chacune de ces thématiques est caractérisée par un ensemble de termes clés qui sont sémantiquement homogènes et discriminants vis-à-vis des termes de l'autre thématique.

La définition ci-dessus désigne une thématique que l'on espère extraire d'un corpus avec des méthodes automatiques. Or, dans la pratique, on obtient souvent du bruit, i.e., des thématiques dépourvues de sens et qui ne représentent aucun intérêt. Afin d'inclure aussi ce type de résultats, nous faisons évoluer la définition formelle d'une thématique vers :

Définition 2 (thématique) *Pour un vocabulaire donné* $\mathbb{V} = \{w_1, w_2, ...w_V\}$, *une thématique* z *est un ensemble de couples* (w_i, p_i) *où* w_i *est un terme du vocabulaire et* p_i *est un score réel positif qui reflète l'importance du terme* w_i *pour la thématique* z. *En d'autres termes :* $z = \{(w_i, p_i), w_i \neq w_j \forall i \neq j \wedge p_i \in \mathbb{R}^+\}$.

Une thématique est généralement décrite par un petit nombre de mots parmi les plus importants en terme du score. Dans ce sens, nous pouvons faire l'analogie avec le problème de classification non supervisée (*clustering*) où une thématique est assimilée à une classe (*cluster*) de termes, celle-ci étant décrite par quelques termes représentatifs. Dans une collection de documents, certaines thématiques sont de tailles importantes (récurrentes) alors que d'autres sont plus rares. Certaines se chevauchent (partagent les mêmes termes), d'autres sont parfaitement séparées.

La modélisation de thématiques est un terme générique qui englobe plusieurs tâches (extraction, nommage, analyse temporelle, évaluation) mais la plus courante est celle de l'extraction. Il s'agit de faire émerger automatiquement les thématiques contenues dans une collection de documents. Cette tâche est parfois accompagnée d'une étape de nommage (donner des noms aux thématiques extraites) ou d'évaluation.

Les différentes problématiques liées à la modélisation de thématiques ont largement été étudiées dans divers contextes comme celui de la fouille de textes, la recherche d'information [56, 127, 128, 129] ou la désambiguïsation lexicale [11, 61]. Cependant, l'application principale de cette tâche demeure l'analyse exploratoire, c'est à dire la découverte de thématiques à partir de grands volumes de données textuelles ("de quoi parle-t-on ?") [39, 80].

Dans le domaine de la recherche d'information, la modélisation de thématiques peut être intégrée dans le processus de recherche d'information de plusieurs manières, par exemple en exploitant la relation entre les documents et les thématiques [129] ou en modélisant les relations entre les thématiques et la requête [56, 128]. Pour la désambiguïsation lexicale, les thématiques sont utilisées comme des contextes supplémentaires sur lesquels se base la sélection du sens d'un mot [11, 61].

Le large spectre d'applications des méthodes d'extraction de thématiques ainsi que la diversité de leur provenance sont deux facteurs qui accentuent la difficulté d'évaluer et de comparer ces méthodes. En effet, ces méthodes produisent des résultats sous différentes formes (distributions de probabilités,

listes de mots, partitions, etc.) qui ont été évalués pour des objectifs différents et dans des contextes différents (recherche d'information, désambiguïsation lexicale, classification non supervisée, etc.). Pour toutes ces raisons, il est difficile de pouvoir comparer des méthodes d'extraction de thématiques dans le même cade et de manière uniforme.

Dans ce travail, nous soulignons les problèmes liées à l'évaluation des méthodes utilisées pour l'extraction de thématiques. Ensuite, nous décrivons une nouvelle mesure, appelée *Vraisemblance Généralisée*, pour évaluer différentes méthodes d'extraction de thématiques dans le même cadre. Cette mesure est basée sur l'extension du concept de la vraisemblance, largement utilisé dans la modélisation probabiliste. Nous montrons la pertinence de la mesure proposée à travers des expérimentations sur des données réelles issues de l'actualité. Le résultat de cette évaluation nous servira par la suite à faire nos choix techniques en ce qui concerne la problématique principale de la thèse (modélisation conjointe des thématiques et des opinions). Ce travail a été publié dans [23].

Ce chapitre est organisé en trois sections : la section 2.2 présente le processus de prétraitement de données textuelles et les méthodes de représentaton. La sction 2.3 présente un état de l'art où nous exposons les problématiques et méthodes liées à la tâche de modélisation de thématiques. Enfin, la section 2.4 décrit notre contribution pour évaluer les méthodes d'extraction de thématiques.

2.2 Prétraitement et représentation de données textuelles

Dans le domaine de la fouille de données, le prétraitement de données est une étape importante dont dépend grandement la qualité des résultats. Généralement, cette étape constitue la plus grande partie du travail en terme de temps de traitement et d'effort humain. Le prétraitement de données consiste non seulement à préparer les données pour l'analyse en les mettant en forme mais aussi à y apporter des transformations [91], par exemple, nettoyage, correction d'erreurs, remplissage de données manquantes, etc. Dans cette section, nous présentons un aperçu du processus du prétraitement des données textuelles.

Le prétraitement de données est une étape systématiquement appliquée avant toute tâche de fouille de données. En effet, la qualité des données en entrée est déterminante pour le déroulement de la tâche et les résultats obtenus. Plusieurs critères déterminent la qualité des données, e.g., l'intégrité, la représentativité du phénomène réel, l'absence d'erreurs et des données incomplètes, etc. Une mauvaise qualité des données résulte très souvent en un résultat biaisé et faussé (*garbage in garbage out*). D'où l'importance de cette

étape de prétraitement qui consiste à mieux préparer les données pour être analysées par la machine. Par ailleurs, un prétraitement efficace devrait aider à optimiser l'utilisation des ressources de la machine en réduisant la taille de l'espace occupé par les données sur le disque ou dans la mémoire.

Le prétraitement de données est à plus forte raison nécessaire quand ces données sont de type textuel. En effet, le texte présente plusieurs particularités qui en font un type de données très vulnérable aux problèmes d'intégrité, de déformation, et d'erreurs, et ce pour plusieurs raisons. D'abord, une grande partie des données textuelles proviennent du Web et sont générés par les internautes (forums, blogs, réseaux sociaux..). Cela impose *a minima* un prétraitement de nettoyage comme la suppression des méta-données par exemple les données du style d'un document HTML. Ensuite, l'hétérogénéité des données textuelles (pages Web, documents de bureautique, PDF) impose un travail supplémentaire d'homogénéisation et d'intégration. Enfin, il est souvent nécessaire de représenter les données textuelles sous une forme exploitable par les algorithmes d'analyse. Nous distinguons généralement deux étapes de préparation de données : prétraitement et représentation.

2.2.1 Prétraitement

La plupart de documents Web sont sous forme semi-structurée, typiquement en HTML où le document contient en plus du texte, des métadonnées de forme et de style. Cela nécessite un prétraitement de nettoyage pour éliminer toutes les données qui ne sont pas du contenu effectif du document. Cette étape convient également pour supprimer les mots vides (*stopwords*) qui n'ont pas ou peu de pouvoir discriminatif comme"un", "avec", "de", etc.

La normalisation est un autre prétraitement qui consiste à réduire les différentes inflexions d'un mot à la même forme, afin de les traiter comme une seule entité. Deux types de normalisation sont généralement utilisées : la lemmatisation (*lemmatization*) et la racinisation (*stemming*). La lemmatisation est un traitement linguistique qui consiste à réduire les mots à leurs lemmes. Les noms sont réduits au masculin singulier, et les verbes à l'infinitif. Par exemple, les mots "efficace", "efficacité", "efficacement" sont transformés en "efficace". La lemmatisation peut éventuellement inclure la classe grammaticale du mot (POS [1]), e.g., nom, verbe, adjectif, comme information supplémentaire. La lemmatisation repose sur le déploiement de ressources linguistiques lourdes comme les analyseurs syntaxiques et les ontologies.

La racinisation est un traitement purement algorithmique qui consiste à supprimer les préfixes et les suffixes des mots et les réduire ainsi à leurs racines. Par exemple, les mots "aimable", "aimer", "aime" sont transformés en "aim". *Porter Stemmer* [90] est un algorithme de racinisation anglophone qui est

1. Part-Of-Speech

devenu très populaire depuis sa proposition en 1980. Il a fait l'objet de multiples extensions afin de couvrir plusieurs langues et de s'intégrer facilement dans plusieurs langages de programmation.

2.2.2 Représentation

La majorité des méthodes de représentation de documents utilisent le modèle vectoriel VSM (*Vector Space Model*) où chaque document est représenté par un vecteur de valeurs numériques ou catégoriques, également appelé sac de mots (*Bag of Word*) [99]. Un corpus (ensemble de documents) est ainsi mis sous la forme d'une matrice documents-termes (cf. figure 2.2.1). Le sens de "terme" ici ne se réduit pas seulement aux mots simples (également appelés uni-grammes) mais inclut aussi d'autres types de segmentation du texte comme les n-grammes de mots ou de caractères. Cette opération est parfois appelée indexation ou segmentation (*tokenization*).

		termes						
		w_1	w_2	w_3	w_4	.	.	w_V
documents	d_1	10	16	1	0	2	4	9
	d_2	15	0	11	0	3	1	0
	d_3	1	12	9	3	5	2	0
	:	2	0	0	13	12	0	10
	d_D	0	2	5	3	13	4	1

X

FIGURE 2.2.1 – Représentation d'un corpus de documents sous la forme d'une matrice X documents-termes avec une pondération TF.

À l'issue de la segmentation, les termes sont caractérisés, chacun, selon son importance par rapport au document dans lequel il apparaît et/ou par rapport aux autres termes du document. Un choix très simple serait d'utiliser la fréquence du terme dans le document TF (*Term Frequency*), mais d'autres critères de pondération existent comme la pondération Booléenne ou encore la pondération TF-IDF (*Term Frequency-Inverse Document Frequency*). Plus récemment, d'autres types de pondération ont été proposés pour optimiser la représentation des documents dans des contextes précise, par exemple la pondération OKAPI utilisée dans un contexte de recherche d'information [96].

La pondération de termes a pour but de caractériser l'importance des termes en se basant sur plusieurs critères (terme fréquent dans le document, terme discriminatif vis-à-vis des autres termes, etc.). Les deux méthodes de pondération Booléenne et TF mesurent l'importance du terme pour le document et pas pour tout le corpus. Cela peut favoriser les termes les plus fréquents, même s'ils sont présents dans beaucoup les documents, par exemple

dans un corpus de critiques de films, des termes comme "film", "acteur" sont fréquents mais pas nécessairement discriminatifs par rapport à une tâche de classification.

TF-IDF est une technique de pondération qui permet de réduire l'effet de ce phénomène. Elle tient compte à la fois de l'importance du terme pour le document et pour le corpus, en pénalisant les termes qui apparaissent dans beaucoup de documents. Dans un corpus de D documents, le poids TF-IDF du terme w_i dans un document d est calculé comme suit :

$$\text{TF-IDF}(w_i, d) = \text{TF}(w_i, d) \times \text{IDF}(w_i) \tag{2.1}$$

Où $\text{IDF}(w_i)$ mesure l'importance du terme w_i pour tout le corpus :

$$\text{IDF}(w_i) = \log(\frac{D}{n_i}) \tag{2.2}$$

Où n_i est le nombre de documents qui contiennent au moins une occurrence du terme w_i.

Le modèle vectoriel transforme le corpus de documents en une matrice directement exploitable par la plupart des méthodes d'analyse. L'utilisation des n-grammes de mots permet de préserver, en partie, l'ordre de mots dans les documents, alors que les n-grammes de caractères donnent plus de robustesse par rapport aux erreurs d'orthographe. En revanche, le modèle vectoriel produit des matrices le plus souvent creuses (avec beaucoup d'éléments nuls), et de dimensions importantes ; en pratique, des dizaines de milliers de termes différents sont extraits à partir d'un corpus de taille moyenne (quelques milliers de documents).

D'autres modèles de représentation existent, comme le modèle de graphe, où les nœuds du graphe représentent les termes du document et les arêtes représentent des relations entre eux (co-occurrence, ordre, etc.). Le modèle de graphe est une représentation qui permet d'exploiter ces données avec des méthodes issues du domaine de la théorie des graphes.

Dans ce manuscrit, nous utilisons une représentation de données basée sur le modèle vectoriel. Ainsi, un corpus de documents est transformé en une matrice documents-termes où les lignes représentent les documents et les colonnes représentent les termes (cf. figure 2.2.1).

2.3 Etat de l'art sur la modélisation de thématiques

Cet état de l'art se décompose en quatre parties correspondant aux trois tâches principales de la modélisation de thématiques, à savoir : l'extraction de thématiques, le nommage et l'évaluation.

2.3.1 Extraction de thématiques

L'extraction de thématiques est une tâche qui consiste à faire émerger automatiquement un ensemble de thématiques à partir d'une collection de documents, le plus souvent de manière non supervisée. La tâche d'extraction de thématiques est à ne pas confondre avec la "segmentation thématique" où il s'agit d'extraire les thématiques conjointement avec les segments de textes qui y sont liés [32].

Dans les dernières années, de nombreux travaux se sont intéressés à cette problématique. En se basant sur le principe de fonctionnement et le type des résultats produits, nous proposons de catégoriser l'ensemble de ces méthodes en trois catégories : modèles probabilistes, méthodes à base de factorisation de matrices et méthodes à base de graphes. Les modèles probabilistes se fondent sur la théorie des probabilités afin d'extraire les thématiques. Les méthodes à base de factorisation de matrices trouvent leurs racines dans l'algèbre linéaire (décomposition de matrices). Enfin, les méthodes à base de graphes exploitent la structure des documents combinée à des techniques de la théorie des graphes. La première approche a été beaucoup plus étudiée que les autres dans la littérature.

Modèles probabilistes

Les modèles probabilistes pour l'extraction de thématiques (*probabilistic topic models*) se basent sur le calcul de probabilités pour l'extraction de thématiques. Ces méthodes s'appuient sur le modèle de Salton [99] où un document est représenté par un vecteur de réels sur l'espace du vocabulaire.

Différentes dénominations. Dans la littérature anglo-saxonne, ces méthodes sont généralement désignées par les termes de *topic models* qui ne fait pas état de la nature probabiliste de ces méthodes. Or, nous pensons qu'il est important que ce mot clé apparaisse dans la dénomination car c'est la caractéristique principale qui différencie cette catégorie. Ces méthodes sont aussi désignées par les termes de "modèles graphiques" car elles peuvent être illustrées avec des graphes à l'instar de beaucoup d'autres méthodes statistiques comme les modèles de Markov cachés, les CRF, etc. Dans ce document, nous utilisons la dénomination "modèles probabilistes".

Principe de base. Le principe de base partagé par l'ensemble de ces méthodes réside dans l'utilisation des lois de probabilité pour caractériser les thématiques. En effet, voici deux hypothèses sur lesquelles se fondent ces méthodes :

- Un document est composé de plusieurs thématiques avec différentes proportions. On parle de "mélange" probabiliste d'un ensemble de thématiques.

– Une thématique est un mélange probabiliste d'un ensemble de termes.

Formellement, pour un vocabulaire $\mathbb{V}$, une thématique z est caractérisée par une distribution de probabilités sur le vocabulaire. Comme le vocabulaire est un ensemble discret de termes, cette distribution est de type multinomiale. Une thématique est représentée par une variable aléatoire z telle que :

$$z \sim \text{Multinomial } (p_1, p_2, ...p_V) \tag{2.3}$$

De même, un document du corpus est représenté par une variable aléatoire d telle que :

$$d \sim \text{Multinomial } (q_1, q_2, ...q_T) \tag{2.4}$$

Idéalement, une thématique doit avoir des probabilités importantes sur un petit nombre de termes et des probabilités faibles sur le reste du vocabulaire. Ces termes sont les termes caractéristiques de la thématique. Par exemple, une thématique relative à la génétique devrait associer des probabilités plus importantes sur les termes "ADN", "gène", "cellule"... que sur les autres.

Modèles graphiques. Un modèle probabiliste peut être représenté par un graphe orienté, d'où la dénomination courante de "modèles graphiques". Les nœuds représentent les différentes variables du modèle (documents, thématiques, termes, etc.) et les arêtes sont des relations de dépendance entre ces variables. Pour illustrer ces propos, nous nous appuyons sur le premier modèle probabiliste pour l'extraction de thématiques PLSA [2] proposé par Hofmann en 1999 [44].

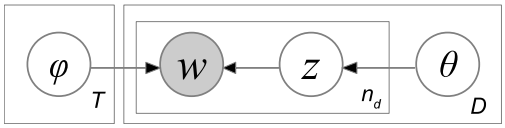


FIGURE 2.3.1 – Modèle graphique de PLSA.

Le modèle graphique de PLSA est représenté par la figure 2.3.1. Ce graphe est lu de la manière suivante :

– Un cadre avec un indice x représente un processus répété x fois.
– Les nœuds grisés sont des variables observées. Tous les autres nœuds (également appelés variables latentes) sont à estimer.
– Le nœud w correspond aux termes des documents.
– Le nœud z correspond aux thématiques.

2. Probabilistic Latent Semantic Analysis/Indexing (plus courant dans le domaine de recherche d'information)

- Le nœud φ représente la distribution multinomiale des thématiques sur le vocabulaire des termes. Il y a T distributions φ_z de ce type à estimer, une pour chaque thématique.
- Le nœud θ représente la distribution multinomiale d'un document sur les thématiques. Il y a D distributions θ_d de ce type à estimer, une pour chaque document.

Processus génératif. Pour un modèle probabiliste donné, il existe un processus génératif qui permet de générer les variables observées quand les autres variables sont connues. Dans le cas de PLSA, il s'agit de générer les mots (et donc les documents) connaissant les distributions φ_z et $\theta_d, \forall z \in \mathbb{Z}, d \in \mathbb{D}$. La génération d'un document d avec PLSA se déroule comme suit :

- Pour chaque terme w à générer :
  1. Tirer une thématique $z \sim \theta_d$.
  2. Tirer un terme $w \sim \varphi_z$.

Ce processus génératif ne permet pas de résoudre complètement la problématique d'extraction de thématiques. Cependant, il demeure très utile car il permet d'illustrer le fonctionnement du modèle et mettre en évidence les différentes dépendances entre les variables. En pratique, le problème consiste à observer les termes des documents et à estimer les distributions φ et θ à partir de ces observations. Ce problème est résolu en utilisant une procédure d'inférence comme par exemple la maximization de vraisemblance (EM) [44], l'inférence variationnelle [9], l'échantillonnage de Gibbs (*Gibbs Sampling*) [39], etc.

Inférence. L'inférence consiste à estimer les paramètres du modèle (variables latentes) à partir des variables observées. Plusieurs méthodes permettent de résoudre ce problème, parmi lesquelles on peut citer :

- Les méthodes basées sur la vraisemblance comme la méthode *Expectation Maximization* (EM) [44], la méthode *Maximum a posteriori* (MAP) [70] et la méthode *Expectation Propagation* (EP) [75].
- Les méthodes à base de l'inférence variationnelle comme dans [9].
- Les méthode de Monte Carlo comme l'échantillonnage de Gibbs [39].

Daud et al. [15] donnent plus de détails sur les différentes méthodes d'inférence avec les modèles probabilistes. Dans ce travail, nous aurons l'occasion de détailler le fonctionnement des deux méthodes : la méthode EM (dans la suite de cette section) et l'échantillonnage de Gibbs (dans la section 4.4.2).

Dans PLSA, l'inférence est réalisée en utilisant la méthode EM. Celle-ci est une méthode itérative où chaque itération se déroule en deux temps : (1) dans l'étape E, les probabilités *a posteriori* $p(z|d, w)$ sont calculées pour chaque variable latente z (thématique) à partir des paramètres courants du modèle ;

(2) dans l'étape M, les paramètres courants du modèle sont mis à jour en se basant sur le résultat de l'étape précédente.

Plusieurs limites de la méthode EM ont été soulignées dans la littérature comme le problème des maxima locaux et la complexité algorithmique due au grand nombre de paramètres à estimer. En outre, suivant la complexité du modèle, le calcul des probabilités *a posteriori* peut être très compliqué, voire impossible. Dans ce cas, plusieurs autres méthodes peuvent être utilisées comme les méthodes basées sur l'inférence variationnelle et les méthodes de Monte Carlo (échantillonnage de Gibbs).

Modèles existants. Outre le modèle PLSA décrit ci-dessus, des dizaines de modèles probabilistes ont été proposés pour l'extraction de thématiques. Parmi ceux-ci, le plus célèbre est sans doute LDA (*Latent Dirichlet Allocation*) [9]. Proposé par Blei et al. en 2003, LDA étend le modèle PLSA avec deux nouvelles variables α et β de type Dirichlet, également appelées hyperparamètres (cf. figure 2.3.2).

Les hyperparamètres de Dirichlet jouent le rôle d'un *a priori* et permettent de contrôler la forme et la densité des distributions multinomiales. Par exemple, des valeurs faibles de α produisent des thématiques plus creuses avec de grandes valeurs de probabilité sur un petit nombre de termes et de faibles valeurs ailleurs. LDA permet aussi de corriger les problèmes de PLSA notamment les problèmes du surapprentissage et de la complexité algorithmique dus au grand nombre de paramètres qu'il faut estimer avec PLSA. Enfin, grâce aux hyperparamètres de Dirichlet, LDA devient un modèle complètement génératif capable d'inférer les thématiques sur de nouveaux documents.

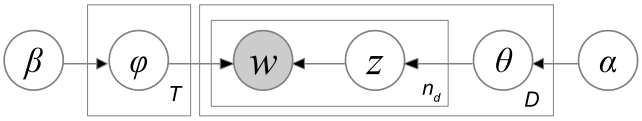


FIGURE 2.3.2 – Modèle graphique de LDA.

Les modèles probabilistes ont l'avantage d'être facilement extensibles. Depuis sa proposition en 2003, le modèle LDA a fait l'objet de dizaines d'extensions. Ces nouveaux modèles qui focalisent sur différents aspects des données textuelles, comme par exemple la corrélation entre les thématiques, le caractère hiérarchique de celles-ci, l'association entre les thématiques et les opinions, l'évolution temporelle, etc. Ces deux derniers aspects seront traités en détails dans ce manuscrit respectivement dans les sections 4.2 et 5.2.

Corrélation entre les thématiques. Plusieurs modèles probabilistes s'intéressent au phénomène de corrélation entre les thématiques. CTM (*Correlated Topic Model*) [6] utilise une matrice de covariance pour mesurer la corrélation

entre chaque paire de thématiques. PAM (*Pachinko Allocation Model*) [62] modélise cette corrélation par un graphe orienté. Dans PAM, une thématique est aussi caractérisée par une distribution sur les autres thématiques. Cette distribution permet de mesurer la corrélation entre les thématiques.

Modèles probabilistes supervisés. Les modèles probabilistes supervisés ont été utilisés pour résoudre des problèmes de classification supervisé ou de régression. sLDA (*supervised LDA*) [8] étend le modèle LDA avec une nouvelle variable de "réponse". Le modèle produit, en plus des thématiques, une réponse pour chaque document qui permet de prédire une classe pour ce document. Dans [55], les auteurs proposent un modèle similaire à sLDA qui prend en entrée une collection de documents annotés. L'objectif est de réaliser une réduction de dimension par la caractérisation des termes discriminatifs en s'appuyant sur l'annotation des documents.

Prise en compte de la structure Les modèles probabilistes ont aussi été déployés pour mixer le contenu et la structure des documents, e.g., fils de discussion, tweets, publications scientifiques, etc. Dans cette veine, ATM (*Author-Topic model*) [97]) permet d'associer les thématiques à des groupes d'auteurs.

Le casse-tête du nombre de thématiques. Dans un domaine comme l'extraction de thématiques, il est important de trouver un nombre "correct" de thématiques T, paramètre obligatoire pour la plupart des méthodes d'extraction de thématiques. Le problème a été principalement étudié pour les modèles de thématiques probabilistes selon deux dimensions : souligner l'importance et l'effet du choix du nombre de thématiques sur les résultats et proposer de nouveaux modèles, dits non-paramétriques, afin d'estimer ce paramètre à partir des données. Dans [2], Arun et al. ont fait remarquer qu'un mauvais choix de ce paramètre dégrade la qualité des résultats de la méthode LDA dans un contexte de classification supervisée par SVM.

Les premiers travaux ont tenté de chercher automatiquement la meilleure valeur de T en optimisant un critère. Dans [9, 97], le nombre de thématiques est fixé en fonction de la perplexité du modèle. Celle-ci est mesurée par la capacité du modèle à générer les données d'un ensemble de test. Une autre méthode consiste à utiliser une sélection de modèles. Une sélection bayésienne est utilisée dans [39] afin de choisir le modèle qui maximise la probabilité *a posteriori* (en intégrant par rapport à toutes les possibilités d'assigner les mots aux thématiques).

D'autres travaux se sont proposés de se passer complètement de ce paramètre. hLDA (*hierarchical LDA*) [5] est un modèle non-paramétrique qui adopte une représentation arborescente avec des thématiques allant de la plus générale à la plus spécifique. Cela évite de fixer le nombre de thématiques *a*

priori mais nécessite en revanche l'interaction de l'utilisateur afin d'exploiter le résultat de l'analyse en spécifiant le niveau de granularité des thématiques. Un travail similaire est proposé dans [121]. Ces modèles non-paramétriques utilisent des méthodes d'inférence adaptées, comme CRP (*Chinese Restaurant Process*) et IBP (*Indian Buffet Process*).

Méthodes à base de factorisation de matrices

En algèbre linéaire, la factorisation (ou décomposition) d'une matrice est sa réécriture comme un produit matriciel de plusieurs matrices. Elle peut être exacte ou approximative. Pour l'extraction de thématiques, le principe général est de partir de la matrice d'occurrences documents-termes X (cf. figure 2.2.1), puis de trouver les matrices W, A, H et E telles que :

$$X = WAH + E. \tag{2.5}$$

H est une matrice dont les lignes sont constituées par des combinaisons de mots caractérisant les thématiques (également appelé espace sémantique latent). W est une matrice de projection des documents dans le cet espace latent. A est une matrice intermédiaire entre W et H qui peut être égale à la matrice identité. Enfin, E est la matrice erreur qui assure l'égalité stricte entre les deux parties de l'équation. Elle peut être égale à la matrice nulle.

Nous avons introduit les deux matrices A et E afin d'unifier les notations et la présentation des principales méthodes de décomposition, celles-ci étant présentées dans la littérature de manière légèrement différente suivant le contexte.

Méthodes existantes. Les premiers travaux d'extraction de thématiques par décomposition de matrices utilisent des techniques mathématiques classiques comme la décomposition en valeurs singulières. Cette méthode est connue sous le nom de LSA (*Latent Semantic Analysis*) [3] [18] car elle transforme la matrice d'occurrences initiale en un espace "sémantique". La décomposition en valeurs singulières peut aussi bien être appliquée sur la matrice X (documents en lignes) que sur sa transposée X^T (documents en colonnes). Ici, nous suivons la deuxième méthode qui est aussi le choix le plus courant dans la littérature. La factorisation se déroule comme suit :

$$X^T = H^T A^T W^T. \tag{2.6}$$

Cette équation produit une factorisation exacte sur un espace latent de dimension $K = \min(N, V)$. Le nombre de thématiques T est généralement choisi

3. Dans le domaine de recherche d'information, la méthode est plus connue sous le nom de *Latent Semantic Indexing*

inférieur à K. La matrice initiale X^T est alors approximée en gardant seulement les T premières valeurs propres de la matrice A^T (cf. figure 2.3.3).

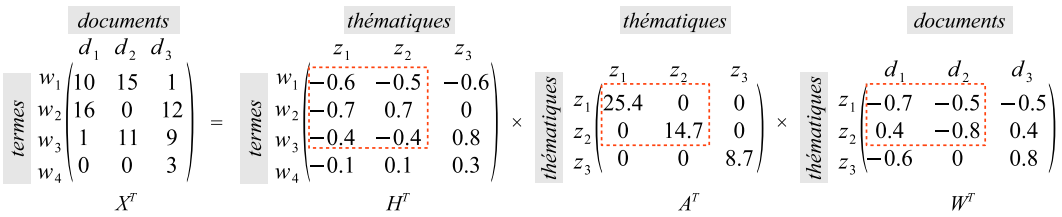


FIGURE 2.3.3 – Une factorisation de matrice avec la méthode LSA. Pour $T = 2$, seulement les deux première valeurs propres sont gardées.

La principale limite de la méthode LSA réside dans la difficulté d'interprétation du résultat et plus particulièrement des vecteurs de la matrice H (thématiques). En effet, LSA mixe des valeurs positives et négatives, ce qui empêche d'interpréter les résultats, notamment dans le contexte d'extraction de thématiques. Certains travaux se sont intéressés à ce point faible de LSA et ont proposé de rendre la factorisation non négative. Cette méthode est connue sous le nom NMF (*Non-negative Matrix Factorization*) [57, 58, 126]. La matrice initiale X est réécrite comme le produit des matrices W et H telles que :

$$X = WH + E, W_{ij} \geq 0 \wedge H_{ij} \geq 0 \ \forall i, j. \tag{2.7}$$

Dans la littérature, il est plus courant d'utiliser une approximation comme l'illustre la figure 2.3.4. La factorisation devient donc :

$$X \approx WH. \tag{2.8}$$

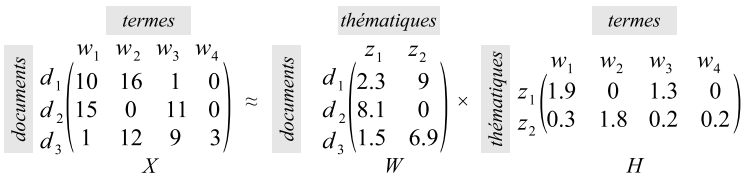


FIGURE 2.3.4 – Une factorisation de matrice avec la méthode NMF basée sur l'algorithme de Lee et Seung [57].

Le problème est donc transformé en un problème d'optimisation sous contraintes qui implique la minimisation de la quantité $||E||$ sous les contraintes de non négativité. Lee et Seung [57] ont proposé de résoudre ce problème en utilisant les multiplicateurs de Lagrange et les conditions de Kuhn-Tucker permettant de résoudre des problèmes d'optimisation sous contraintes d'inégalité

non-linéaires. Cette méthode a donné lieu à un algorithme itératif dont chaque itération comporte les deux opérations suivantes :

$$W \leftarrow W\frac{XH^T}{WHH^T},\ H \leftarrow H\frac{W^TX}{W^TWH}. \tag{2.9}$$

En plus d'être le premier travail dans ce domaine, cet algorithme est aussi le plus connu. D'autres variantes existent comme l'algorithme de Lee et Seung [58] où les règles de mise à jour des matrices W et H sont additives.

Autres méthodes

Méthodes à base de distance. Les méthodes à base de distance se fondent sur le calcul d'une distance pour mesurer la similarité entre les documents. Plusieurs types de distance peuvent être utilisées à partir du moment où les documents sont représentés par des vecteurs dans l'espace du vocabulaire, e.g., distance euclidienne, Cosinus, etc. La plupart des méthodes de cette catégorie sont des méthodes de classification automatique non supervisée. Même si ce n'est pas leur vocation initiale, ces méthodes peuvent être utilisées pour l'extraction de thématiques en considérant que chaque classe forme une thématique et regroupe ainsi les documents qui y sont liés. La caractérisation des thématiques peut ensuite se faire en post-traitement, en prenant par exemple les termes fréquents dans chaque classe ou une combinaison linéaire sur l'espace des termes, e.g., le centroïde.

Dans les méthodes à base de distance, on retrouve principalement les méthodes de partitionnement et les méthodes hiérarchiques. Les méthodes dites de partitionnement, comme les k-Means, commencent par répartir les documents sur un certain nombre de classes et, à chaque itération, les documents sont re-distribués de telle sorte que chacun soit dans la classe dont il est le plus proche (au sens de la mesure de similarité utilisée). La méthode FCM (*Fuzzy c-Means*) est une variante des k-Means qui permet une classification floue [27], i.e. qu'un document peut être lié à plusieurs classes avec différents degrés. Quelques résultats de l'utilisation de ce type de méthodes pour l'extraction de thématiques sont donnés dans [24, 80].

Les méthodes hiérarchiques procèdent à la construction des classes au fur et à mesure par agglomération ou par division. En agglomération, chaque classe contient au départ un seul document. Les deux classes les plus proches, en termes de distance, sont ensuite fusionnées récursivement jusqu'à ce que tous les documents soient dans la même classe. En division, tous les documents sont dans une seule classe qui est divisée récursivement pour l'obtention de classes plus fines.

Dans [12], une méthode agglomérative a été utilisée pour la classification de documents. Dans [88], une méthode hiérarchique est proposée pour la classification de documents en se basant sur une distance qui prend en compte les

entités temporelles et les noms de lieux.

Les méthodes de classification hiérarchiques offrent la possibilité de contrôler la granularité des classes, et d'avoir ainsi des classes aussi fines ou grandes que souhaité. En revanche, les méthodes hiérarchiques sont généralement d'une complexité importante, ce qui les rend inadaptées aux grands volumes de documents.

Qu'elles soient à base de partitionnement ou hiérarchique, ces méthodes n'ont pas comme objectif premier l'extraction de thématiques. Cependant, dans la plupart des cas, un simple post-traitement permet d'y parvenir, dans le sens où les centroïdes correspondent à des vecteurs dans l'espace du vocabulaire de termes. Cela explique la présence de ces méthodes dans cette étude.

Méthodes à base de graphes. Les méthodes à base de graphe reposent sur l'existence d'un ensemble de graphes, orientés ou non, qui représentent la collection de documents. Il s'agit par exemple d'un graphe où les documents sont reliés avec des liens de type "hypertexte". Plusieurs travaux proposent de représenter les pages Web de cette manière, e.g., [35, 53]. Le graphe est ensuite exploité en utilisant des techniques de segmentation afin d'extraire les thématiques, elles aussi représentées par des graphes.

Outre l'avantage de pouvoir s'adapter à de grands volumes de données grâce notamment aux techniques "locales" de segmentation de graphes [35], les méthodes à base de graphes aident à mieux visualiser et interpréter les thématiques et les relations entre elles.

2.3.2 Nommage des thématiques

La phase d'extraction de thématiques est souvent suivie d'une phase de nommage où les thématiques sont décrites par des noms par exemple un ensemble de termes ou une phrase.

Définition 3 (Nommage des thématiques) *Le problème de nommage des thématiques équivaut à trouver une fonction f_{nom} qui associe à chaque thématique un nom. Celui-ci peut prendre différentes formes, comme par exemple un ensemble de termes, une phrase, une liste ordonnée de termes, etc.*

La plupart des méthodes de nommage utilisent les termes fréquents [1, 9, 44]. D'autres travaux essayent de résoudre le problème de nommage en le ramenant à un problème d'apprentissage automatique [130] ou un problème d'optimisation [71].

L'évaluation des noms fournis par les méthodes de nommage est une tâche importante, d'autant plus que les noms sont de formes différentes (mots, ensembles de mots, phrases, etc.). Selon [71, 120, 130], trois critères sont essentiels pour évaluer la qualité des noms :

- Concision : le nom donné à la thématique doit être le plus court possible, mais toutefois suffisant pour décrire la thématique
- Compréhensibilité : la relation entre le nom et la thématique à laquelle il est associé doit être claire : "pourquoi ce nom pour ces documents ?"
- Caractère distinctif : les termes qui constituent le nom doivent apparaître dans les thématiques qu'ils représentent et pas, ou rarement, dans les autres.

Dans [71], le problème de nommage est transformé en un problème d'optimisation impliquant la maximisation de la divergence KL entre les noms candidats et les thématiques. La procédure générale se déroule comme suit :

1. Un ensemble de noms candidats est généré pour chacune des thématiques extraites de manière linguistique en s'appuyant sur des analyseurs syntaxiques pour générer des phrases et des expressions et/ou de manière statistique en utilisant des n-grammes.
2. Les noms candidats sont ensuite ordonnés sur la base de similarité sémantique avec la thématique. Celle-ci est calculée pou un nom b et une thématique z de la manière suivante :

$$\mathrm{SIM}(b, z) = \sum_{w \in P} p(w|z) \cdot \mathrm{PMI}(w, b|\mathcal{C}) - \mathrm{KL}(z||\mathcal{C}) \tag{2.10}$$

   où PMI est l'information mutuelle (*Point-wise Mutual Information*) et $\mathcal{C}$ représente un contexte qui peut être construit en prenant une grande quantité de données relatives au domaine.

D'autres méthodes de nommage ont été proposées comme le nommage par les motifs (*itemsets*) fréquents [34], les mots clés [118] et les concepts [47, 67]. Fung et al. [34] se sont proposés de traiter le problème de mots rares en filtrant les motifs non fréquents. Wartena et Brussee [118] ont utilisé des mots clés qui sont généralement plus informatifs et spécifiques au domaine. Les méthodes de nommage avec les concepts sont généralement basées sur des indicateurs statistiques pour la création des concepts en les combinant avec les ontologies et les bases de connaissances par exemple *Google Directory Service* [67] et DBpedia [47].

2.3.3 Evaluation

Les méthodes d'extraction de thématiques ne sont pas toutes évaluées de la même manière. Les premiers travaux dans l'extraction de thématiques ont été développés pour la recherche d'information. La supériorité d'une méthode par rapport à une autre se mesurait par sa capacité à fournir des résultats plus précis dans un contexte de recherche d'information, e.g., précision, rappel, F-score. Avec l'avènement des modèles probabilistes et la diversification des

tâches associées à l'extraction de thématiques, plusieurs autres évaluations ont été utilisées, comme la perplexité qui est égale à la probabilité de générer les données connaissant les paramètres du modèle. La problématique d'évaluation des thématiques se retrouve aussi en apprentissage non supervisé avec les mesures recensées par Halkidi et al. [40]. Le résultat d'extraction de thématiques peut ainsi être évalué en le rapprochant d'un résultat de classification non supervisée.

Différentes catégorisations. Les mesures d'évaluation existant dans ce domaine peuvent être catégorisées de deux manières différentes :

- **Mesures qualitatives vs. mesures quantitatives** : l'approche qualitative a recours au jugement humain pour qualifier les thématiques et elle est généralement non quantifiée. A contrario, l'approche quantitative permet de mesurer de manière quantifiée et plus précise la qualité des modèles, qu'elle soit basée sur le jugement humain ou non. L'une des mesures quantitatives qui n'utilisent pas le jugement humain (automatiques) est la vraisemblance mais elle est seulement calculable pour les modèles probabilistes, ce qui n'est pas le cas de toutes les méthodes d'extraction de thématiques.
- **Mesures externes vs. mesures internes** : les mesures externes évaluent la qualité des résultats par rapport à une référence définie par les classes des documents fixées *a priori*. Comme exemples de ce type de mesures on peut citer le F-score (moyenne harmonique du rappel et de la précision), l'entropie (mesure de désordre dans l'ensemble des thématiques) et la pureté (ratio moyen de la classe majoritaire dans chacune des thématiques). En revanche, les mesures internes ne font pas appel à des connaissances extérieures. Par exemple, l'inertie intra-classe utilisée comme fonction objectif dans la méthode des k-Means ou la cohésion [105] qui mesure la similarité Cosinus entre les documents d'une même thématique.

Un problème de spécificité. Comme il a été montré ci-dessus, chacune des mesures est spécifique à un problème particulier. Les mesures issues de la recherche d'information ne s'avèrent pas suffisantes dans un contexte où l'on veut évaluer l'homogénéité des thématiques. Les mesures à base de vraisemblance (par exemple la perplexité) ne sont calculables que pour les modèles probabilistes. Les mesures issues du domaine de la classification non supervisée ne prennent pas en compte la spécificité des données textuelles (par exemple le phénomène de chevauchement entre les thématiques). Enfin, les mesures externes posent des problème d'application pratique car la vérité terrain est souvent une information coûteuse à obtenir et éminemment subjective.

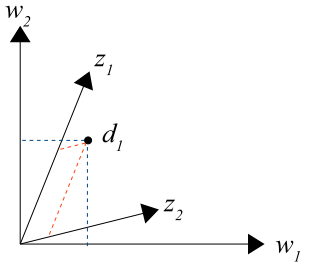
Un problème d'hétérogénéité. Les méthodes d'extraction de thématiques produisent des résultats de formes hétérogènes car elles proviennent de domaines différents. Afin de décrire les thématiques, les modèles probabilistes produisent des distributions de probabilités alors que les méthodes à base de factorisation de matrices produisent des matrices. Cela pose clairement un obstacle majeur pour évaluer toutes ces méthodes dans le même cadre et de manière uniforme. A notre connaissance, il n'existe pas à ce jour de mesure automatique qui permet d'évaluer toutes les méthodes présentées de manière uniforme, et ainsi de pouvoir les comparer.

2.4 Contribution : évaluation des méthodes d'extraction de thématiques

Comme il a été montré dans l'état de l'art, les différentes méthodes d'extraction de thématiques produisent des résultats de forme hétérogène : partitions de documents, distributions de probabilités sur les mots, matrices, etc. Cela pose un problème pour évaluer et comparer ces méthodes dans le même cadre. Notre contribution à la résolution de ce problème consiste à proposer un nouvel espace de description commun aux différentes méthodes et une nouvelle mesure de qualité qui se calcule dans cet espace. Ainsi, il devient possible de comparer des méthodes de nature différente et ce de manière quantitative.

La mesure que nous proposons, intitulée *Vraisemblance Généralisée* (VG), est une mesure quantitative interne qui permet d'évaluer plusieurs méthodes d'extraction de thématiques, même si ces dernières sont basées sur des modèles mathématiques différents. Le bien-fondé de la mesure VG repose sur le fait qu'il existe une analogie entre les différentes méthodes. En effet, toutes ces méthodes permettent, d'une manière ou d'une autre, de projeter les documents dans un espace de description formé par les thématiques et de décrire ces thématiques par une combinaison linéaire du vocabulaire. La figure 2.4.1 illustre ces propos. Le document d_1 par exemple est plongé dans l'espace formé par les deux thématiques z_1 et z_2 que nous appelons "espace latent". Notre contribution consiste à définir des transformations des résultats des méthodes vers l'espace latent ainsi que la mesure VG qui se calcule dans cet espace latent.

La mesure VG est calculée à partir de deux matrices : la matrice de projection W, matrice de projection des documents dans l'espace latent et la matrice de l'espace latent H qui définit cet espace. La matrice H est caractérisée par un ensemble de vecteurs décrits dans l'espace des termes (vocabulaire) (cf. figure 2.4.1). Ce sont donc des vecteurs non négatifs. La matrice W est caractérisée par un ensemble de vecteurs correspondant aux documents, décrits dans l'espace des thématiques. En d'autres termes, les matrices W et H sont telles que :



$$\begin{array}{c} \\ d_1 \\ d_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} w_1 & w_2 \end{array} \\ \begin{pmatrix} 10 & 16 \\ 15 & 4 \end{pmatrix} \end{array} \approx \begin{array}{c} \\ d_1 \\ d_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} z_1 & z_2 \end{array} \\ \begin{pmatrix} 3 & 1 \\ 0 & 4 \end{pmatrix} \end{array} \times \begin{array}{c} \\ z_1 \\ z_2 \end{array}\!\!\begin{array}{c} \begin{array}{cc} w_1 & w_2 \end{array} \\ \begin{pmatrix} 2 & 5 \\ 4 & 1 \end{pmatrix} \end{array}$$

$X \qquad\qquad W \qquad\qquad H$

FIGURE 2.4.1 – Espace latent : les documents sont projetés dans l'espace latent caractérisé par les thématiques z_1 et z_2 (matrice W) et les thématiques sont décrites par une combinaison linéaire de termes (matrice H).

- W_{ij} est un score mesurant la relation entre le document d_i et la thématique z_j.
- H_{jk} est un score mesurant la relation entre la thématique z_j et le terme w_k.

Nous définissons trois transformations vers l'espace latent pour trois méthodes issues des principales approches présentées dans la section 2.3.1 : LDA, NMF et FCM. Pour la méthode LDA, W_{ij} est la probabilité $p(d_i|z_j)$, et H_{jk} est la probabilité $p(w_k|z_j)$. Pour la méthode NMF, les deux matrices sont directement obtenues par factorisation. Pour la méthode FCM, W_{ij} correspond au degré d'appartenance du document d_i à la classe z_j, et le vecteur H_{j*} correspond au centroïde de la classe z_j.

Les deux matrices W et H doivent être normalisées (si elles ne l'étaient pas déjà) afin d'avoir un même ordre de grandeur quelque soit la méthode utilisée et d'éviter ainsi un biais éventuel dans le calcul de la mesure :

- $\sum_{j=1}^{Z} W_{ij} = 1, \forall i \in \{1..D\}$ (normalisation des lignes de W).
- $\sum_{k=1}^{V} H_{jk} = 1, \forall j \in \{1..T\}$ (normalisation des lignes de H).

Sous ces hypothèses, nous définissons score(d_i, w_k), le score de vraisemblance d'une occurrence du terme w_k dans le document d_i comme suit :

$$\text{score}(d_i, w_k) = \sum_{j=1}^{T} W_{ij} \times H_{jk} \tag{2.11}$$

En d'autres termes, score(d_i, w_k) est obtenu en multipliant la ligne de la matrice W qui correspond au document d_i par la colonne de la matrice H qui correspond au terme w_k.

Le score de vraisemblance d'un document d_i est défini comme suit :

$$\text{score}(d_i) = \prod_{w \in \mathbb{V}} \text{score}(d_i, w)^{n(d_i, w)} \tag{2.12}$$

Où $n(d_i, w)$ est le nombre d'occurrences du mot w dans le document d_i. En passant au log :

$$\log \text{score}(d_i) = \sum_{w \in \mathbb{V}} n(d_i, w) \log \text{score}(d_i, w) \tag{2.13}$$

La mesure VG est basée sur la moyenne géométrique des scores individuels des documents, score(d_i), chaque score étant lui-même un produit calculé sur chaque terme du vocabulaire (cf. équation 2.12). La multiplication géométrique a donc la forme d'un produit de produits.

Pour normaliser, il suffit de mettre à la puissance inverse du nombre de facteurs dans la multiplication. Ce nombre est égal à la double somme : $\sum_{d_i \in \mathbb{D}} \sum_{w \in \mathbb{V}} n(d_i, w)$. Au final, la mesure VG est calculée, pour un corpus $\mathbb{D}$, avec la formule suivante :

$$\text{VG}(\mathbb{D}) = \exp \left\{ \frac{\sum_{d_i \in D} \log \text{score}(d_i)}{\sum_{d_i \in \mathbb{D}} \sum_{w \in \mathbb{V}} n(d_i, w)} \right\} \tag{2.14}$$

La mesure VG est largement inspirée de la mesure de perplexité utilisée pour évaluer les modèles probabilistes [114]. Celle-ci est basée sur le concept de la vraisemblance que l'on peut interpréter par la capacité du modèle à générer un ensemble de documents. En réalité, la perplexité est l'inverse de la moyenne géométrique des vraisemblances calculées pour les documents séparément. La mesure VG est l'extension de la perplexité pour couvrir d'autres méthodes non nécessairement probabilistes. Nous nous sommes basés sur les mêmes formules utilisées pour le calcul de la perplexité. En revanche, nous avons choisi d'utiliser simplement la moyenne géométrique de la vraisemblance au lieu de l'inverse de celle-ci. Par conséquent, la mesure VG est à maximiser.

La mesure VG peut être calculée sur un corpus de test différent du corpus sur lequel les thématiques ont été extraites (corpus d'apprentissage) mais ceci suppose que le modèle soit prédictif, c'est-à-dire capable d'affecter les nouveaux documents de test aux thématiques déjà extraites. Ceci n'est malheureusement pas le cas de toutes les méthodes, notamment les méthodes d'apprentissage non supervisé et les méthodes à base de factorisation de matrices. Pour cette raison et afin de mieux interpréter le résultat de la mesure VG, nous avons choisi de l'évaluer sur la base du corpus d'apprentissage.

2.5 Expérimentations

Dans cette section, nous présentons le protocole expérimental (corpus, prétraitements, outils, paramètres des méthodes, etc.), ainsi que les résultats obtenus.

Corpus	AP	Elections
Langue	Anglais	Français
Nombre de documents (D)	2210	2777
Taille du vocabulaire (V)	9794	9855

TABLE 2.1 – Présentation des corpus AP et Elections.

Protocole expérimental. Les tests sont effectués sur deux corpus : AP et Elections. AP est un corpus de documents de l'agence de presse *Associated Press* [42], également utilisé dans [71, 78]. Elections est un corpus de documents Web (médias, blogs, réseaux sociaux, etc.), qui traitent des élections présidentielles françaises de 2012. Ces documents ont été collectés durant la période du 16/03/2012 au 16/04/2012 par la plateforme de veille AMIEI (cf. section 6.2). Le tableau 2.1 résume le contenu de chaque corpus, après les prétraitements suivants :

- Suppression de mots outils, par exemple "le", "sur", "dans", etc.
- Racinisation (*stemming*), par exemple les mots "logement" et "loger" deviennent "log"
- Suppression des mots qui occurrent une seule fois dans le document.

Les tests sont réalisés pour les méthodes LDA, NMF et FCM. Afin de limiter le risque de tomber dans des optima locaux, le même test est réalisé 5 fois et la moyenne est retenue. Les paramètres de la méthode LDA sont fixés sur la base des règles généralement utilisées dans la littérature [39] : $\alpha = \frac{50}{T}, \beta = 0.01$, nombre d'itérations = 1000. Les paramètres de la méthode FCM sont fixés empiriquement : $m = 1.1$, nombre maximum d'itérations = 20. Pour exécuter LDA, nous nous sommes appuyés sur l'outil Mallet [69]. Pour NMF, nous avons utilisé notre propre implémentation et, pour FCM, nous avons utilisé le langage R [92].

Les deux types d'expérimentations réalisés sont les suivants :

- Test de comparaison : les trois méthodes LDA, NMF et FCM sont comparées suivant les scores obtenues par la mesure VG.
- Tests sur les cas extrêmes : deux cas extrêmes sont considérés : *Crisp* (chaque document est lié à une seule thématique à la fois) et *Uniforme* (chaque document est lié à toutes les thématiques avec le même score). Les résultats correspondant à ces deux cas extrêmes sont créés artificiellement en fixant le score à 1 pour la thématique qui maximise le score obtenu par NMF dans le cas *Crisp* et en fixant le score à $\frac{1}{T}$ pour toutes les thématiques dans le cas *Uniforme*.

Résultats. Un extrait des résultats d'exécution des trois méthodes LDA, NMF et FCM sont représentés dans le tableau 2.2. Nous remarquons sur ce tableau que les thématiques extraites par LDA et NMF sont facilement interpétables, contrairement à celles extraites par la méthode FCM. Cela laisse

Thématiques	**immobilier**	**économie**	**vote des étrangers**	**sondages**
LDA	commun	économie	droit	sondage
	prix	crise	valeur	erreur
	immobilier	payer	vote	marge
	paris	argument	immigrés	institut
	politique	dire	étrangers	candidat
	crédit	marché	politique	journal
NMF	logement	euro	droit	sondage
	immobilier	politique	étrangers	institut
	construction	dollar	local	marge
	encadrement	part	municipal	erreur
	loyer	marché	gauche	harris
	prix	économie	non	ifop
FCM	spatial	page	fou	situation
	riom	divorce	marketing	copain
	municipal	seul	basculer	crever
	défaite	immatriculation	fier	croissance
	ajaccio	identifier	tronquer	sage
	bras	dollar	démontrer	attirer

TABLE 2.2 – Exemple de thématiques découvertes par les trois méthodes sur le corpus Elections ($T = 50$). Les noms ont été donnés manuellement.

présager que ce résultat est de qualité inférieure par rapport à celui obtenu avec les méthodes LDA et NMF.

La comparaison des trois méthodes par la mesure VG est représentée dans la figure 2.5.1. Ce test montre que les méthodes LDA et NMF présentent un comportement similaire au vu de la variation de la mesure VG en fonction du nombre de thématiques. En effet, cette dernière augmente avec l'augmentation du nombre de thématiques. Ce phénomène peut être expliqué par le fait qu'un trop petit nombre de thématiques mène à mélanger plusieurs thématiques dans une seule et donne ainsi des résultats de moindre qualité. En revanche, un grand nombre de thématiques permet de mieux séparer les thématiques, permet aux thématiques de petite taille d'émerger et donne ainsi un résultat de meilleure qualité. Nous remarquons que la mesure VG suit parfaitement cette logique.

Cependant, nous remarquons que si le nombre de thématiques est encore plus grand (proche du nombre de documents), les résultats convergent vers un modèle où une thématique est extraite pour chaque document. La valeur de la mesure VG continue à augmenter sans pour autant que le résultat soit forcément de meilleure qualité. Ce problème est similaire au problème de surapprentissage (*overfitting*) connu dans le domaine de l'apprentissage statistique.

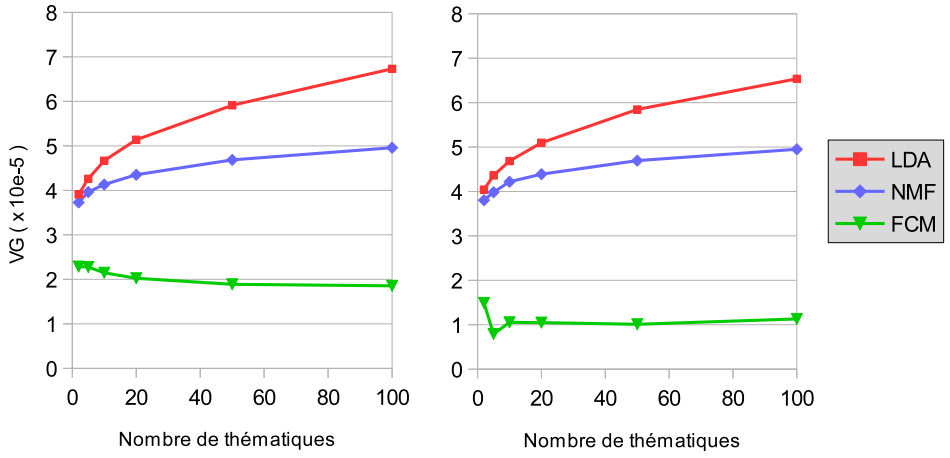


FIGURE 2.5.1 – Variation de la mesure VG (à maximiser) en fonction du nombre de thématiques sur le corpus AP (gauche) et Elections (droite).

LDA demeure la méthode qui donne les meilleurs résultats, en termes de la mesure VG, par rapport à NMF et FCM, et ce sur les deux corpus (cf. figure 2.5.1). La qualité des résultats donnés par la méthode FCM est remarquablement inférieure en termes de la mesure VG à celle des deux autres méthodes. Ceci est conforme à l'évaluation qualitative illustrée sur le tableau 2.2. En effet, les thématiques extraites par la méthode FCM sont mélangées et très difficiles à interpréter.

Les résultats obtenus pour les cas extrêmes sont représentées dans la figure 2.5.2. L'objectif de ce test est d'analyser le comportement de la mesure VG dans les deux cas extrêmes : *Crisp* et *Uniforme* (cf. section 2.5). Les résultats obtenus avec ces deux cas de figure sont censés être de moins bonne qualité par rapport aux résultats obtenus avec LDA et NMF. En effet, ces deux cas représentent des configurations où un document est lié à une seule thématique ou alors à toutes les thématiques, ce qui est intuitivement loin de la réalité. En effet, un document est généralement lié à un petit nombre de thématiques.

Suivant VG, les deux cas extrêmes *Crisp* et *Uniforme* sont des configurations moins bonnes que celle produite par NMF (cf. figure 2.5.2). Cela confirme que ces deux cas de figure ne donnent pas de bons résultats et qu'un bon ensemble de thématiques constitue en général un compromis entre les deux extrêmes, à savoir quelques thématiques pertinentes pour un document.

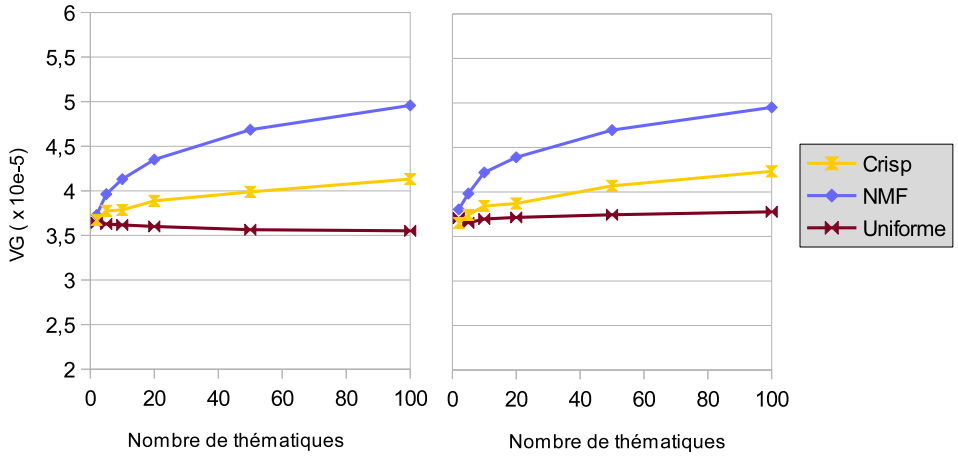


FIGURE 2.5.2 – Variation de la mesure VG (à maximiser) en fonction du nombre de thématiques dans les cas extrêmes sur les corpus AP (gauche) et Elections (droite).

2.6 Discussion

Les méthodes d'extraction de thématiques, étant issues de domaines variés, produisent des résultats de forme hétérogène, ce qui empêche leur comparaison de manière uniforme. Nous avons proposé une mesure d'évaluation, la Vraisemblance Généralisée, qui permet d'évaluer dans un cadre commun les méthodes d'extraction de thématiques. Pour calculer la mesure, les résultats de ces méthodes sont transformés dans un nouvel espace qui plonge les documents dans l'espace latent des thématiques.

La mesure de la Vraisemblance Généralisée que nous avons proposée a permis de comparer trois méthodes d'extraction de thématiques (LDA, NMF et FCM) sur deux corpus différents. Les résultats ont donné l'avantage à la méthode LDA, suivie de NMF puis de FCM. Les résultats donnés par la méthode d'apprentissage FCM étaient d'une qualité inférieure, selon la mesure VG, par rapport aux deux autres méthodes. Ceci a été conforme avec une analyse qualitative des thématiques extraites par cette méthode. En effet, ces dernières étaient mélangées et très difficiles à interpréter.

Cependant, l'utilisation de la vraisemblance comme unique critère d'évaluation, comme dans le cas de la mesure VG, risque d'introduire un biais. En effet, certaines méthodes sont basées sur l'optimisation de cette entité afin d'extraire les thématiques, par exemple les modèle probabilistes utilisant l'algorithme EM comme algorithme d'inférence. Dans ce cas, la vraisemblance des données

d'apprentissage est implicitement optimisée à chaque itération de l'algorithme. L'analyse de présence de biais éventuels dans notre évaluation nécessite des investigations plus poussées que nous avons laissé pour un futur travail.

Il serait également intéressant, en complément à ce travail, de tester le comportement de la mesure VG sur des corpus de test (différents des corpus d'apprentissage). Cela nécessiterait la définition des opérations de prédiction pour les méthodes d'extraction de thématiques afin de pouvoir affecter les nouveaux documents aux thématiques.

Chapitre 3

Modélisation d'Opinions

Résumé. *Dans ce chapitre, nous nous intéressons à la tâche de l'analyse d'opinions. Nous exposons les principales approches existantes pour la détection et le classement de l'opinion à partir des données textuelles, en mettant l'accent sur la problématique d'analyse de polarités. A travers des exemples concrets, nous illustrons les limites des approches basées sur l'apprentissage automatique supervisé, notamment la méthode Naive Bayes notamment le problème de sur-apprentissage, et nous proposons une méthode hybride (combinant l'apprentissage supervisé et la connaissance a priori) afin d'améliorer les résultats du classement.*

Sommaire

3.1 Introduction

L'opinion est une composante de la pensée humaine qui peut être définie comme étant un point de vue ou une position prise par rapport à un objet, un service, une idée, un événement.. ce que l'on peut appeler la cible de l'opinion. Elle peut être exprimée de différentes manières (expressions, gestes, etc.). Selon le dictionnaire Larousse, une opinion est un "jugement, avis, sentiment qu'un individu ou un groupe émet sur un sujet, des faits, ce qu'il en pense". L'opinion peut donc prendre différentes modalités : une opinion positive, négative, neutre ; une opinion de soutien, de détraction ; une opinion de doute, de peur, de contentement et ainsi de suite.

Les données textuelles sur le Web peuvent être objectives ou subjectives. Les données objectives expriment des faits alors que les données subjectives expriment des opinions qui émanent de leurs créateurs. L'analyse d'opinions, également appelée fouille d'opinions, est une tâche qui consiste à détecter et classer les données textuelles exprimant ce type d'information subjective. Généralement, la problème de classement fait s'opposer des classes d'opinions divergentes comme par exemple objective/subjective, positive/négative/neutre, joie/peur/dégoût. Dans ce dernier cas, on parle préférentiellement d'analyse d'émotions (*emotion analysis*) [28].

Avec l'avènement du Web 2.0, les quantités de données sur le Web, potentiellement subjectives, ne cessent de se multiplier. La communication de pair à pair rendue possible par ces technologies par exemple les forums de discussion, les blogs et les réseaux sociaux, permet à tout internaute d'être producteur et consommateur de l'information. Ce mode de communication est un facteur principal de la croissance démesurée des données subjectives.

Cette abondance de données subjectives de nature variée est d'une grande utilité, autant pour les producteurs que pour les consommateurs de l'information. Les premiers travaux en analyse d'opinions utilisaient les données issues des sites de critique (chambres d'hôtels, restaurants, films, ..) [10, 51, 86]. Les avis laissés par les utilisateurs ont été analysés, classés et résumés, et ce à des fins exploratoires, par exemple l'intelligence économique, la e-réputation, ou prédictifs par exemple la recommandation. Plus récemment, beaucoup de travaux se sont tournés vers les réseaux sociaux, à l'instar de Twitter qui a été utilisé pour prédire les ventes au cinéma [3] ou encore les sondages politiques [82].

Dans les années 2000, la problématique d'analyse d'opinions a été traitée par des chercheurs qui venaient de domaines variés (informatique, linguistique, psychologie, sociologie). L'analyse d'opinions couvre généralement plusieurs tâches : détection de textes subjectifs, classement de textes positifs/négatifs/

neutres, modélisation de l'intensité (plutôt positif/positif/très positif), etc. Cependant, la littérature en informatique met plus en avant le problème de classement de textes positifs/négatifs/neutres (analyse de la polarité).

Dans ce chapitre, nous nous intéressons au domaine de l'analyse de polarités d'opinions où beaucoup de travaux sont basés sur l'apprentissage automatique supervisé. Nous montrons, à travers des expérimentations et des exemples, les principales limites de ce type de méthodes en particulier la méthode *Naive Bayes*. En effet, cette méthode s'avère particulièrement sensible à la qualité de données d'apprentissage notamment quand ces données sont déséquilibrées ou sous-représentatives (phénomène connu sous le nom de sur-apprentissage). Afin d'atténuer ce problème, nous proposons une nouvelle méthode qui combine l'apprentissage automatique et la connaissance *a priori* exprimée sous forme d'un lexique d'opinions. Nous démontrons sur différents jeux de données l'apport de notre méthode pour limiter les effets dus au sur-apprentissage en améliorant les scores de classement de l'opinion. Ces travaux ont été publiés dans [20].

La première section de ce chapitre présente un bref état de l'art de l'analyse d'opinions. La section 3.3 présente notre contribution. Les sections 3.4 et 3.5 présentent respectivement les expérimentations et la discussion des résultats.

3.2 Etat de l'art

L'analyse d'opinions est un domaine qui réunit plusieurs problématiques autour de la détection et le classement de données subjectives :

- Analyse de subjectivité : déterminer si un mot, une phrase ou un texte est subjectif (exprime une opinion) ou objectif (exprime des faits).
- Analyse de polarité : déterminer si un mot, une phrase ou un texte exprime une opinion positive, négative ou neutre.
- Analyse d'intensité : déterminer la polarité d'un mot, d'une phrase ou d'un document mais avec une granularité plus fine (très positif, positif, plutôt positif, plutôt négatif, négatif, très négatif).
- Analyse d'émotions : l'émotion est un concept plus large que l'opinion. La tâche consiste à classer les textes selon plusieurs modalités d'émotion. Par exemple : peur, joie, tristesse, surprise, etc.

Depuis le début des années 2000, beaucoup de travaux se sont intéressés aux différentes problématiques liées à l'analyse d'opinions. Cependant, une grande partie de ces travaux se sont focalisés sur l'analyse de polarité. Depuis quelques années, l'intérêt s'est porté sur les données issues de microblogs et de réseaux sociaux. L'apparition de ces nouveaux modes de communication a motivé les travaux autour de ces types de données et a donné lieu à plusieurs

campagnes d'évaluation internationales comme par exemple SemEval [1], CLEF Initiative [2], SemSA (ESWC'2014) [3], etc.

L'analyse d'opinions couvre un large spectre de travaux provenant de différentes communautés (apprentissage automatique, TAL, psychologie, sciences cognitives, etc.). Dans cette section, nous présentons brièvement les travaux significatifs dans le domaine d'analyse d'opinions en les catégorisant en méthodes d'apprentissage automatique, méthodes à base de règles et méthodes à base de similarité entre les termes. Cet état de l'art se focalise principalement sur les travaux de l'analyse de polarité.

3.2.1 Méthodes d'apprentissage automatique supervisé

Pour l'analyse d'opinions, les méthodes d'apprentissage automatique supervisé construisent un modèle sur la base d'un corpus annoté, i.e., une collection de documents, chacun associé à une classe d'opinion. Le modèle est ensuite déployé pour classer de nouveaux documents.

L'analyse d'opinion avec des méthodes d'apprentissage supervisé équivaut essentiellement à se ramener à un problème d'apprentissage et d'y appliquer les méthodes classiques. Les données textuelles sont représentées dans une matrice d'occurrences documents-termes [99] et la classe (variable endogène) correspond à la polarité de l'opinion. Il faut choisir la méthode d'apprentissage qui appréhende le mieux les particularités des données textuelles (données creuses, dimensionnalité importante..) ainsi qu'au choix des variables explicatives (termes). De plus amples détails sur les modèles de représentation existants et les configurations de termes/pondérations ainsi que les méthodes de sélection de variables ont été donnés dans la section 2.2.

Plusieurs méthodes d'apprentissage automatique supervisé ont été utilisées pour le classement d'opinions. Il est difficile de donner ici un état de l'art exhaustif. Nous proposons donc de citer quelques travaux significatifs et d'en détailler le fonctionnement.

NB. *Naive Bayes* (NB) est une méthode basée sur l'utilisation de probabilités conditionnelles. Soit c une variable qui désigne la classe d'opinion d'un document. Le principe général est de calculer, à partir des documents d'apprentissage, toutes les probabilités de termes connaissant la classe d'opinion :

$$p(w|c) = \frac{1}{\text{nb}(c)} \cdot \text{nb}(w, c) \tag{3.1}$$

1. http ://alt.qcri.org/semeval2015/
2. http ://www.clef-initiative.eu/
3. http ://challenges.2014.eswc-conferences.org/index.php/SemSA

où nb(c) est le nombre de documents de classe c et nb(w, c) est le nombre de documents de classe c qui contiennent le terme w.

Une fois ces probabilités calculées pour chaque couple (w, c), le modèle peut être utilisé pour classer un nouveau document d en calculant les probabilités $p(c|d)$. En utilisant le théorème de Bayes et la règle d'indépendance entre les termes, nous obtenos :

$$p(c|d) \propto p(c) \cdot \prod_{w \in d} p(w|c) \tag{3.2}$$

Le document est généralement affecté à la classe qui maximise cette probabilité suivant l'approche MAP (*Maximum A Posteriori*).

SVM. *Support Vector Machines* (SVM) est une autre approche de classement qui a largement été utilisée pour l'analyse d'opinions et pour le classement binaire de documents de manière plus générale. Le principe est de trouver une fonction qui projette les données (documents) dans un espace de grande dimension. La fonction doit permettre de maximiser l'hyperplan entre les documents de classes opposées et par conséquent réduire les erreurs de classement. SVM est une méthode qui se prête bien aux données textuelles pour sa capacité à gérer les données creuses.

Pang et al. [86] ont expérimenté les méthodes d'apprentissage NB et SVM ainsi qu'une méthode à base d'entropie (MaxEnt) pour l'analyse d'opinions. Ils ont combiné plusieurs types de pondération de termes (Booléen, TF, TF-IDF) et plusieurs types de termes (mots, n-grammes, POS, adjectifs). Les meilleurs scores ont été obtenus par SVM en utilisant une pondération booléenne (présence/absence des termes).

Malgré la popularité des méthodes d'apprentissage automatique, celles-ci présentent plusieurs limites qui empêchent un déploiement efficace. Ces méthodes nécessitent une phase d'apprentissage et des données annotées, i.e., un ensemble de documents où chacun est étiqueté avec une classe d'opinions, qui doit être suffisant (en terme de quantité) et représentatif (de bonne qualité). Les modèles appris de cette manière reflètent les structures contenues dans les données d'apprentissage. Il devient donc important de veiller à la bonne qualité des données d'apprentissage et de limiter toute forme de biais ou de déséquilibre.

3.2.2 Méthodes à base de règles

L'analyse d'opinions a également été approchée sous un angle linguistique, en utilisant des règles de classement par exemple des règles syntaxiques et lexicales. Les méthodes à base de règles n'utilisant pas forcément la représentation vectorielle puisque l'ordre et la séquentialité des mots deviennent des critères

importants pour décider de la subjectivité et de la polarité. Certaines de ces méthodes utilisent aussi l'apprentissage automatique en le combinant avec des traitements linguistiques *a priori*. On parle dans ce cas de méthodes hybrides. Nous reviendrons sur ces travaux avec plus de détails dans la section 3.3.2.

Les méthodes à base de règles reposent sur l'existence d'un ensemble de règles linguistiques généralement définies à la main pour détecter les signaux subjectifs dans le texte. Chacune de ces règles permet de détecter un mot, une structure syntaxique ou alors une expression et apporte ainsi une information qui sera utilisée pour le classement d'opinions.

#	Règle	Détection de	Type	Signification
1	{‘‘bon’’, ‘‘meilleur’’}	polarité	mot	un mot parmi {“bon”, “meilleur”}
2	‘‘aim*’’	polarité	expression régulière	mots commençant par “aim” (e.g., “aimer”, “aimable”, etc.)
3	(ADV, ADJ)	subjectivité	grammatical	un adverbe suivi par un adjectif
4	(NEG, ADJ)	subjectivité	négation	négation d' adjectifs
5	(NEG ‘‘trouver’’ POS_PHRASE)	polarité	grammatical	une phrase positive précédée par la négation du verbe “trouver”
6	ONT.SYN (‘‘mauvais’’)	polarité	ontologie	tous les synonymes du mot “mauvais” présents dans l'ontologie ONT

Table 3.1 – Exemples de règles linguistiques pour l'analyse d'opinions.

Les règles linguistiques diffèrent par leur forme et leur complexité. Le tableau 3.1 présente quelques exemples de règles linguistiques allant des plus simples (présence de certains mots) aux plus compliquées (structures grammaticales et utilisation d'ontologies). Dans ce tableau, les règles sont représentées en utilisant une notation mathématique. La règle #1 permet de détecter les signaux positifs en se basant sur la présence des mots “bon” et “meilleur”. La règle #2 permet de détecter la présence de tous les mots qui commencent par “aim”. Ce type de règles est utile pour réduire l'espace de recherche en détectant toutes les inflexions d'un mot en une seule fois. La règle #3 permet de détecter les adverbes suivis d'un adjectif. La règle #4 permet de détecter la négation d'un adjectif (changement de polarité). La règle #5 permet de chercher le patron correspondant à la négation du verbe “trouver” suivie d'une

phrase positive. Enfin, la règle #6 utilise l'ontologie ONT afin de trouver tous les synonymes du verbe "mauvais".

Un grand nombre de règles est généralement utilisé pour l'analyse d'opinions. Celles-ci peuvent aussi être combinées pour générer des règles plus complexes, par exemple les règles #3 et #4 ci-dessus, et/ou pondérées afin de réaliser une analyse plus fine. Dans la littérature, beaucoup de travaux dans cette catégorie proposent aussi la détection de la cible d'opinion (*opinion target*). Nous reviendrons sur ces travaux avec plus de détails dans la section 4.2.1.

Les travaux présentés dans [68] constituent un bon exemple représentatif de ce type de méthodes. Les auteurs ont défini manuellement un ensemble de règles (ici appelées grammaires) pour la détection de polarité à partir de textes francophones. Trois grammaires ont été développées pour être appliquées à trois domaines différents (tourisme, livres, jeux vidéos). Le texte est transformé en une liste de "relations". Par exemple, le phrase "je n'aime pas aller au cinéma" est transformée en `SENTIMENT_NEGATIF(aimer, aller)`. Ces relations sont ensuite moyennées pour déduire la polarité globale du texte.

Kennedy et Inkpen [51] ont proposé une méthode basée sur le travail de [111]. Celui-ci a été étendu par un ensemble de règles syntaxiques afin de capturer le changement dans l'intensité pour l'Anglais. Ils ont examiné l'effet de trois types d'intensité "very", "deeply", "rather". Pour cela, un ensemble de règles a été défini manuellement pour :

- identifier des termes positifs/négatifs sur la base d'un lexique d'opinion enrichi de [111],
- identifier la négation et les adverbes qui augmentent/diminuent l'intensité des adjectifs,
- réaliser une désambiguïsation lexicale.

Une fois ces informations extraites, les documents sont convertis en une représentation vectorielle avant d'appliquer la méthode de classement SVM. Un résultat légèrement meilleur a été obtenu par rapport à l'utilisation d'un modèle vectoriel avec les mots simples.

Wilson et al. [123] ont proposé un travail similaire pour la détection de subjectivité au niveau de la phrase. La démarche générale se déroule en deux étapes : (1) détection des phrases neutres, et (2) analyse de polarité des phrases subjectives. Pour la détection du neutre, la phrase est caractérisée par les attributs suivants :

- Termes (mots simples, mots du lexique, POS)
- Attributs de modification (présence d'adverbes)
- Attributs de la phrase (nombre d'adjectifs, de pronoms, ..)
- Attributs de structure (voix active/passive)
- La thématique du document (cf. chapitre 2).

Pour l'analyse de polarité, les phrases subjectives sont caractérisées par des attributs de polarité (présence de modificateurs de polarité, de négation, ..).

Dans [81], les auteurs ont examiné l'apport de certains types d'attributs linguistiques, comme les n-grammes et les termes d'un lexique d'opinions, dans deux contextes différents : identification de critiques (analyse de subjectivité) et analyse de polarité. Les scores obtenus pour l'analyse de subjectivité étaient de loin meilleurs par rapport à l'analyse de polarité. Cela montre que le problème d'identification de textes subjectifs peut être résolu en utilisant ce type d'attributs, notamment le lexique de termes d'opinions, contrairement à l'analyse de polarité qui s'avère une tâche plus complexe.

Les méthodes à base de règles permettent une analyse fine du langage et ainsi une meilleure gestion de certains aspects de la langue, comme la négation, les phrases d'opposition (e.g., "mais"), co-référence, etc. Cependant, la création de ces règles peut devenir une tâche laborieuse. D'abord, parce que ces règles sont très dépendantes du domaine. Ensuite, la définition des règles nécessite un travail manuel et une bonne connaissance du domaine. Enfin, le traitement linguistique de grandes quantités de données nécessite le déploiement d'outils linguistiques complexes, ce qui rend ces méthodes peu adaptées aux grands volumes de données.

Il faut noter en conclusion que les règles de classement peuvent servir à construire de nouveaux attributs pour les méthodes d'apprentissage automatique [17, 26, 100, 101, 124]. Cela permet alors de mixer la connaissance provenant des données et celle provenant des experts. On parle quelquefois de méthodes hybrides. Nous reviendrons sur ces méthodes avec plus de détails dans la section 3.3.2.

3.2.3 Méthodes à base de similarité entre termes

Dans cette section, nous nous intéressons aux méthodes qui utilisent la similarité entre les termes comme base de fonctionnement. Ces méthodes prennent initialement une liste réduite de mots positifs et négatifs, appelée noyau (*seed list*), et procèdent à l'enrichissement de cette liste avec de nouveaux mots qui sont "proches". La phase d'enrichissement est basée sur le calcul d'une similarité entre les mots du noyau et les mots du corpus. Celle-ci peut être statistique comme la corrélation et la co-occurrence ou linguistique comme les relations de synonymie et d'antonymie.

Turney et Littman [111] ont proposé un noyau de 14 mots positifs et négatifs pour l'Anglais, choisis car ils ne sont pas spécifiques à un domaine :

Positifs = { *good, nice, excellent, positive, fortunate, correct, superior* }
Négatifs = { *bad, nasty, poor, negative, unfortunate, wrong, inferior* }

Une similarité sémantique entre deux mots, un mot w_1 du noyau et un mot w_2 du corpus est calculée de deux manières différentes : PMI (*Point-wise Mutual*

Information)et LSA (*Latent Semantic Analysis*), toutes les deux basées sur les co-occurrences de mots. Soit $p(w)$ la proportion de documents contenant le mot w. Soit $p(w_1\text{ET } w_2)$ la proportion de documents contenant à la fois les deux mots w_1 et w_2. La similarité PMI entre deux mots w_1 et w_2 est calculée comme suit :

$$\text{PMI}(w_1, w_2) = \log \frac{p(w_1\text{ET } w_2)}{p(w_1) \times p(w_2)} \tag{3.3}$$

Où $p(w)$ mesure la fréquence d'apparition d'un mot w dans le corpus.

La similarité LSA se calcule à partir de la factorisation de la matrice documents-termes (cf. section 2.3.1) :

$$\text{LSA}(w_1, w_2) = \cos(v_1, v_2) \tag{3.4}$$

où v_1, v_2 sont les vecteurs correspondant à la projection des mots w_1, w_2 dans l'espace latent. Enfin, pour une mesure de similarité, notée SIM, la polarité d'un mot w est calculée comme suit :

$$\text{Polarité}(w) = \sum_{w_i \in \text{Positifs}} \text{SIM } (w, w_i) - \sum_{w_j \in \text{Negatifs}} \text{SIM } (w, w_j) \tag{3.5}$$

La polarité du mot w correspond au signe de la différence calculée ci-dessus.

Kamps et al. [50] ont proposé une méthode où la similarité entre deux mots est mesurée en se basant sur WordNet [74]. Un graphe de synonymes est construit en considérant comme synonymes tous les mots qui se trouvent dans le même groupe lexical (appelé *synset*) de WordNet. Ces synonymes ainsi reliés par des arêtes constituent un graphe. Ensuite, la similarité entre deux mots de ce graphe est mesurée par la longueur du plus court chemin entre eux. Les auteurs ont défini trois types de noyaux :

- le noyau {*good*, *bad*} pour l'analyse de polarité,
- le noyau {*strong*, *weak*} pour l'analyse d'intensité,
- le noyau {*active*, *passive*} pour l'analyse de l'activité.

Harb et al. [41] ont proposé de résoudre le problème de spécificité de certains mots du noyau qui peuvent être dépendants du domaine, i.e., pour un domaine donné, un mot peut ne pas exister ou avoir un sens différent. Les auteurs ont utilisé un corpus d'apprentissage pour chaque domaine, obtenu avec un moteur de recherche. Ensuite, un ensemble d'adjectifs positifs/négatifs est construit à partir de chaque corpus d'apprentissage en se basant sur la liste noyau de [111], laquelle est enrichie avec de nouveaux adjectifs (à partir du corpus). Cette phase d'enrichissement est basée sur le calcul de la corrélation entre les adjectifs du corpus et les adjectifs de la liste noyau. Enfin, un texte est classé selon le nombre d'adjectifs positifs/négatifs qu'il contient.

Les méthodes à base de similarité entre les mots ont permis la création automatique de nombreuses ressources lexicales, comme SentiWordNet [30] et le lexique de Pak et Paroubek pour le Français [83]. D'autres ressources ont été créés manuellement avec des annotations plus fines, comme le lexique MPQA [123] et le lexique de B. Liu [25].

3.2.4 Evaluation

L'analyse d'opinions étant souvent assimilée à une tâche de classification supervisée, elle peut être évaluée par les mesures d'évaluation recensées par Halkidi et al. [40]. Le plus simple serait de considérer la proportion de documents qui sont correctement classés par le système (taux de succès). Dans la littérature, les méthodes d'analyse d'opinions sont généralement évaluées sur la base des mesures du rappel et de la précision. Ces deux mesures sont calculées pour chacune des classes d'opinions séparément avant d'être moyennées sur toutes les classes. Le calcul des mesures de rappel et de précision pour une classe c se fait en considérant toutes les autres classes comme opposées à la classe c. Toutes les autres classes sont alors notées $\bar{c}$. Ensuite, le tableau de validation croisée est construit en calculant quatre types de fréquences (voir tableau 3.2) : les vrais positifs (VP), les faux positifs (FP), les faux négatifs (FN) et les vrais négatifs (VN). Enfin, les mesures sont calculées à partir du tableau de validation croisée.

		Prédiction	
		c	$\bar{c}$
Réalité	c	VP	FN
	$\bar{c}$	FP	VN

Table 3.2 – Calcul des mesures de rappel et de la précision

Le rappel mesure la capacité du système à donner tous les résultats pertinents. Il est calculé pour la classe c comme suit :

$$\mathrm{R}(c) = \frac{\mathrm{VP}}{\mathrm{VP+FN}} \tag{3.6}$$

La précision mesure la capacité du système à refuser tous les résultats non pertinents. Elle est calculée pour la classe c comme suit :

$$\mathrm{P}(c) = \frac{\mathrm{VP}}{\mathrm{VP+FP}} \tag{3.7}$$

Enfin, la mesure F-score est calculée pour la classe c comme la moyenne harmonique du rappel et de la précision. Elle mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres.

$$\text{F-score}(c) = 2 \times \frac{\text{R}(c) \times \text{P}(c)}{\text{R}(c) + \text{P}(c)} \tag{3.8}$$

La mesure d'évaluation globale est la moyenne arithmétique de toutes les mesures de F-score calculées individuellement sur chacune des classes. Deux méthodes existent pour calculer cette moyenne : macro F-score et micro F-score. Dans la macro F-score, le même poids est donné à toutes les classes quelles que soient leurs tailles. Cela est réalisé en calculant une moyenne arithmétique simple. Dans la micro F-score, les scores calculés individuellement pour les classes sont pondérés par les tailles de celles-ci. Dans ce manuscript, nous utilisons la macro F-score car c'est la méthode la plus utilisée dans ce domaine.

3.3 Contribution : une méthode hybride pour l'analyse d'opinions

S'appuyant uniquement sur les données en minimisant l'interaction humaine, les méthodes d'apprentissage supervisé constituent une bonne alternative aux méthodes qui se basent sur une analyse linguistique fine. Dans ce travail, nous nous intéressons à la méthode d'apprentissage statistique NB, connue pour sa popularité dans le domaine d'analyse d'opinions du fait de ses résultats comparables à d'autres méthodes, comme SVM et MaxEnt, et de sa faible complexité algorithmique par rapport aux autres méthodes. De plus, NB est basée sur des théories de probabilités relativement plus simples à comprendre et à implémenter.

Cependant, la méthode NB peut s'avérer fortement dépendante de la qualité des données d'apprentissage. Dans cette section, nous soulignons les limites de la méthode NB dans le domaine d'analyse d'opinions sur le Web. En particulier, nous montrons que cette méthode a tendance à sur-apprendre et produit parfois des modèles très dépendants des données d'apprentissage et peu généralisables sur de nouvelles données. Afin d'atténuer ce problème, nous proposons une nouvelle approche qui consiste à incorporer une connaissance *a priori* durant la phase d'apprentissage de NB. Cette connaissance est représentée sous forme d'un lexique d'opinion générique et indépendant de tout domaine. Nous montrons, en nous appuyant sur des données Web réelles, que notre approche améliore les performances de NB et reste compétitive avec SVM tout en préservant une faible complexité algorithmique.

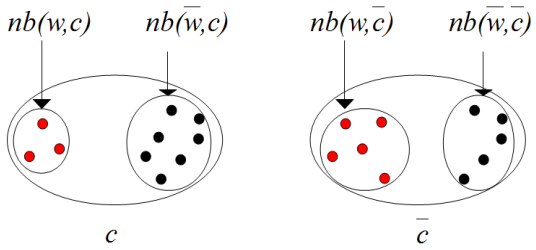


FIGURE 3.3.1 – Illustration de la méthode NB. En rouge, les documents contenant le terme w. En noir, les documents ne contenant pas le terme w.

3.3.1 Limites de la méthode d'apprentissage NB

Comme nous l'avons vu précédemment dans la section 3.2.1, la méthode NB est basée sur le calcul de probabilités de termes connaissant les classes $p(w|c)$ comme suit :

$$p(w|c) = \frac{1}{\text{nb}(c)} \cdot \text{nb}(w, c) \tag{3.9}$$

où $\text{nb}(c)$ est le nombre de documents de classe c et $\text{nb}(w, c)$ est le nombre de documents de classe c qui contiennent le terme w (voir figure 3.3.1).

Cette information est ensuite stockée et utilisée pour la prédiction des classes sur de nouveaux documents comme suit :

$$p(c|d) \propto p(c) \cdot \prod_{w \in d} p(w|c) \tag{3.10}$$

La qualité des données d'apprentissage est déterminante pour les performances et la généricité du modèle. Or, ces données, héritant de la nature des données Web, présentent plusieurs problèmes potentiels (biais, déséquilibre, bruit, etc). Des travaux ont déjà souligné la sensibilité de la méthode NB aux données déséquilibrées [33, 93]. A cela, nous rajoutons le problème de biais éventuel dans les données d'apprentissage qui peut être lié à différents facteurs, comme la collecte, l'échantillonnage, etc.

Déséquilibre. Le déséquilibre concerne la différence d'effectif (taille) entre les classes. Comme il a déjà été souligné dans [33, 93], le modèle NB apprend moins bien sur les classes de faibles effectifs, ce qui rend l'affectation des documents à la classe majoritaire plus probable et génère des erreurs. Dans [93], les auteurs ont proposé *Complement*-NB, un modèle similaire à NB qui procède par apprentissage sur une classe en l'opposant à toutes les autres classes, de manière itérative. Cette technique a permis d'améliorer les résultats dans un contexte de classement thématique.

Dans le contexte d'analyse d'opinions, le problème de déséquilibre est présent. Des expérimentations ont montré que le modèle NB a un mauvais score de rappel sur les classes de petites tailles. Nous reviendrons sur les détails de ces expérimentations dans la section 3.3.2, tableau 3.11.

Biais. Le biais peut être engendré par la méthode de collecte de données, par exemple la restriction sur un ensemble de mots clés, de pages Web ou d'utilisateurs. Cela donne une certaine spécificité aux données d'apprentissage, ce qui n'est pas nécessairement le cas des données de test ou de généralisation.

La deuxième forme de biais est située au niveau des termes. Suivant le domaine, certains termes peuvent apparaître plus fréquemment dans une classe que dans une autre sans que ces mots soient forcément spécifiques à cette classe. Cela peut conduire à des erreurs de classement. Par exemple, si le mot "acteur" apparaît plus dans un contexte négatif, il sera classé comme négatif par le modèle NB alors qu'il ne l'est pas.

Afin d'illustrer ces propos, nous avons réalisé des expérimentations sur le corpus de tweets de la campagne d'évaluation SemEval'2013, tâche 2 [122]. SemEval est une campagne d'évaluation annuelle pour un ensemble de tâches linguistiques et/ou statistiques. La tâche 2 de la campagne SemEval'2013 consistait à classer les opinions dans des données issues de Twitter. Les systèmes participants ont effectué leur apprentissage sur un ensemble de données composé de 10404 tweets manuellement annotés par la polarité d'opinion positive, négative ou neutre (notées c_+, c_- et c_0 respectivement). Les systèmes étaient ensuite évalués sur la base d'un ensemble de test composé de 3813 tweets annotés de la même manière. Le tableau 3.3 donne quelques statistiques sur ce corpus.

Echantillon	Taille	$\|c_+\|$	$\|c_-\|$	$\|c_0\|$
Apprentissage	10404	3859	1634	4911
Test	3813	1572	601	1640

TABLE 3.3 – Corpus de tweets SemEval [122].

Nous nous sommes appuyés sur le corpus d'apprentissage SemEval pour créer un modèle NB. Le tableau 3.4 montre des exemples de termes en situation de biais, i.e., qui sont plus fréquents dans la polarité contraire à leur polarité *a priori*. Par exemple, le mot "*mad*" exprime *a priori* une polarité négative alors qu'il est plus fréquent dans les documents annotés positivement. L'avant-dernière colonne du tableau montre des exemples de tweets que la méthode NB n'a pas réussi à classer correctement à cause du biais présent dans les données d'apprentissage.

Terme (w)	Polarité *a priori*	Proba. donn. appr. ($\times 10^{-2}$)		Exemples d'erreurs de classement NB (c_r=classe réelle, c_p=classe prédite)		
		$p(w\|c_+)$	$p(w\|c_-)$	tweet	c_r	c_p
honesty	POS	0.03	0.18	There is an honesty that says we play the game Saturday..	c_+	c_-
critical	NEG	0.08	0.00	Need cough medicine and/or Lemsips. May be critical if you want the Gazette on Wednesday !..	c_-	c_+
fault	NEG	0.08	0.06	no matter how you slam Obama, you own our credit rating downgrade. It is ALL your fault..	c_-	c_+
mad	NEG	2.23	1.96	Tomorrow is KARWA- CHOTH Delhi is going MAD !!!!!!!..	c_-	c_+

TABLE 3.4 – Exemples de termes en situation de biais pour un modèle de classement NB (corpus de tweets SemEval). Les colonnes 2 et 3 montrent la fréquence des termes dans les tweets de classes positive et négative.

3.3.2 Notre proposition : une méthode hybride

Afin d'atténuer les effets de déséquilibre et de biais présentés ci-dessus, nous proposons une nouvelle approche basée sur NB mais qui prend en compte une connaissance *a priori*. Cette connaissance est représentée sous forme d'un lexique d'opinions (liste de termes positifs et négatifs). Cette connaissance *a priori* permet au modèle NB d'être moins dépendant des données d'apprentissage et de mieux appréhender les nouvelles données.

Travaux similaires. L'idée d'incorporer une connaissance *a priori* au sein d'un modèle d'apprentissage automatique n'est pas nouvelle dans le domaine de classification de données textuelles [17, 26, 100, 101, 124]. Dans [100], le modèle de régression logistique a été modifié dans ce sens afin de prendre en compte les données d'apprentissage au même titre qu'un ensemble de pseudo-données générées et annotées manuellement. Dans [17], les paramètres (mode, variance) d'un ensemble de termes annotés manuellement ont été amplifiés afin que ces termes puissent contribuer davantage au classement des documents. D'autres techniques pour incorporer de la connaissance ont été proposées avec la méthode SVM [124] et les réseaux de neurones [101].

Le travail le plus proche du nôtre est celui de Melville et al. [72]. Leur modèle prend en compte deux types de probabilités : celles apprises par le modèle NB et celles définies par des experts. Ces probabilités sont ensuite agrégées afin de construire un modèle de classement hybride. Notre approche est basée sur un principe similaire d'adaptation de probabilités mais nous adoptons une méthode qui consiste à intégrer une connaissance extérieure dans le modèle de classement.

Terme	score_{c_+}	score_{c_-}
mad	0	1
good	1	0
:-)	1	0
not so good	0	1
...		

TABLE 3.5 – Représentation de la connaissance *a priori* par un lexique de termes polarisés.

Notre méthode. La connaissance *a priori* est représentée sous forme d'un lexique de termes où chaque terme est annoté avec des scores d'appartenance aux deux polarités de l'opinion : positive et négative (cf. tableau 3.5). Cette approche peut tout à fait être généralisée pour un problème de classement à trois polarités (positive, négative et neutre).

L'idée générale consiste à adapter les valeurs de probabilités $p(w|c)$ calculées uniquement à partir des données de telle sorte à prendre en compte cette connaissance. Prenons l'exemple du terme *mad*. Les fréquences de ce teme dans le corpus SemEval ont déjà été donnés dans la tableau 3.4. Les probabilités NB correspondant à ces fréquences sont :

$$\begin{aligned} p(\text{“}mad\text{”}|c_+) &= \frac{86}{3859} \\ &= 0.0223. \\ p(\text{“}mad\text{”}|c_-) &= \frac{32}{1634} \\ &= 0.0196. \end{aligned} \tag{3.11}$$

Notre idée consiste à combiner ces probabilités avec celles données par les experts (lexique) afin de prendre en compte à la fois la connaissance provenant des données d'apprentissage et celle provenant des experts. Pour ce faire, nous avons expérimenté plusieurs stratégies et nous en avons retenu deux pour leur bons résultats : *Add & Remove* et *Transfer*. Ces deux stratégies diffèrent dans le calcul des probabilités $p(w|c)$ mais utilisent la même règle de classement basée sur le maximum *a posteriori* : $\text{classe}(d) = \arg\max_c p(c|d)$.

Add & Remove. Soit c la classe *a priori* d'un terme w du lexique (déterminée par le lexique). Cette stratégie consiste à rajouter, artificiellement, des occurrences du terme w du lexique à la classe c et d'en supprimer de l'autre classe $\bar{c}$. Cela contribue à rééquilibrer la distribution du terme w sur les polarités d'opinion en la rapprochant de la connaissance expert définie par le lexique.

Par conséquent, cette stratégie aide à corriger les erreurs de classement dûes à la présence de biais et de déséquilibre.

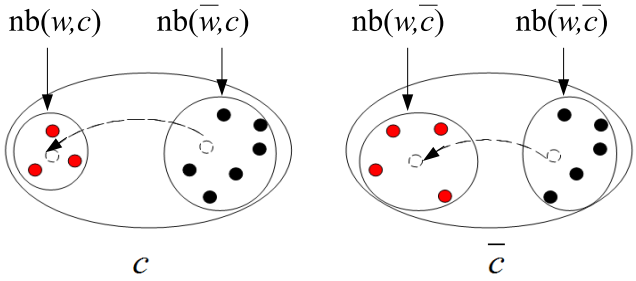


FIGURE 3.3.2 – Illustration de la méthode Add&Remove. En rouge, les documents contenant le terme w. En noir, les documents ne contenant pas le terme w.

Pour s'assurer que les valeurs des probabilités ne dépassent pas 1, nous introduisons $\text{nb}(\bar{w}, c)$, le nombre de documents de la classe c qui ne contiennent pas le terme w et qui représente le nombre maximum d'occurrences du terme w qui peuvent être rajoutées à la classe c (voir figure 3.3.2). Ainsi, le nombre d'occurrences effectivement rajoutées est un ratio α_c de ce maximum ($0 \leq \alpha_c \leq 1$).

De même, si c n'était pas la classe *a priori* du terme w, le nombre d'occurrences qui sont supprimées de la classe c est un ratio β_c du nombre maximum qui peut être supprimé $\text{nb}(w, c)$, avec $0 \leq \beta_c \leq 1$. Formellement, les probabilités sont calculées comme suit :

$$\begin{aligned} p(w|c) &= \frac{1}{\text{nb}(c)} \cdot [\text{nb}(w,c) + \alpha_c \cdot score_c(w) \cdot \text{nb}(\bar{w},c) \\ &\quad -\beta_c \cdot score_{\bar{c}}(w) \cdot \text{nb}(w,c)] \end{aligned} \tag{3.12}$$

Reprenons l'exemple du terme *mad* se trouvant en situation de biais. Avec les paramètres $\alpha_+ = \alpha_- = 0.002, \beta_+ = \beta_- = 0.3$ communs aux deux classes d'opinions, nous avons les nouvelles valeurs des probabilités suivantes :

$$\begin{aligned} p(\text{“}mad\text{”}|c_+) &= \frac{86 + 0 - 0.3 \times 1 \times 86}{3859} \\ &= 0.0156. \\ p(\text{“}mad\text{”}|c_-) &= \frac{32 + 0.002 \times 1 \times 3773 - 0}{1634} \\ &= 0.0242. \end{aligned} \tag{3.13}$$

Transfer. Cette stratégie consiste à transférer des occurrences d'un terme w vers sa classe *a priori* à partir de l'autre classe. Le nombre d'occurrences transférées est tel qu'il ne soit pas plus grand que le nombre effectif d'occurrences dans la classe d'origine et que la probabilité finale ne soit pas supérieure à 1. Soit $max(w, c)$ le nombre maximum d'occurrences du terme w qui peuvent être transférées d'une classe c à l'autre classe $\bar{c}$ (voir figure 3.3.3). Ce nombre ne doit pas être plus grand que le nombre de documents de $\bar{c}$ contenant w ni au nombre de documents de c ne contenant pas w.

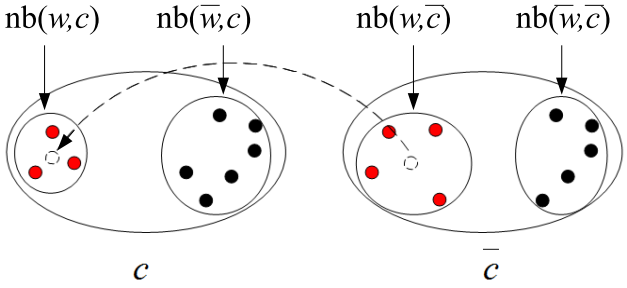


FIGURE 3.3.3 – Illustration de la méthode Add&Remove. En rouge, les documents contenant le terme w. En noir, les documents ne contenant pas le terme w.

$$max(w, c) = \min\{\text{nb}(w, \bar{c}), \text{nb}(\bar{w}, c)\} \tag{3.14}$$

Enfin, le nombre d'occurrences effectivement transférées est un ratio α_c de $max(w, c)$ avec $0 \leq \alpha_c \leq 1$. Les probabilités finales sont calculées comme suit :

$$\begin{aligned} p(w|c) &= \frac{1}{\text{nb}(c)} \cdot [\text{nb}(w, c) + \alpha_c \cdot score_c(w) \cdot max(w, c) \\ &\quad -\alpha_c \cdot score_{\bar{c}}(w) \cdot max(w, \bar{c})] \end{aligned} \tag{3.15}$$

Pour les paramètres $\alpha_+ = \alpha_- = 0.3$, le terme *mad* en situation de biais a les valeurs de probabilités suivantes :

$$\begin{aligned} p(\text{“}mad\text{”}|c_+) &= \frac{86 + 0 - 0.3 \times 1 \times 32}{3859} \\ &= 0.0120. \\ p(\text{“}mad\text{”}|c_-) &= \frac{32 + 0.3 \times 1 \times 32 - 0}{1634} \\ &= 0.0137. \end{aligned} \tag{3.16}$$

Les deux stratégies, *Add & Remove* et *Transfer* reviennent à rajouter des occurrences des termes du lexique à la classe *a priori* et d'en supprimer de

l'autre classe. Dans la méthode *Transfer*, le nombre d'occurrences rajoutées est exactement égal au nombre d'occurrences supprimées.

Méthode	$p(\text{“}mad\text{”}\|c_+)$	$p(\text{“}mad\text{”}\|c_-)$
NB	0.0223	0.0196
Add&Remove	0.0156	0.0242
Transfer	0.0120	0.0137

TABLE 3.6 – Exemple récapitulatif des probabilités obtenues par les trois méthodes NB, Add&Remove et Transfer pour le terme *mad* à partir du corpus SemEval.

Exemple récapitulatif. Le tableau 3.6 récapitule le calcul de probabilités pour le terme *mad* à partir du corpus de tweets SemEval. Nous remarquons sur ce tableau que la méthode NB associe à ce terme une probabilité plus importante sur la polarité positive, et ce conformément à sa distribution dans les données d'apprentissage. Cette configuration n'est pas bonne car ce terme est *a priori* négatif (selon le lexique), ce qui conduit à des erreurs de classement avec la méthode NB. Avec notre méthode représentée par les deux stratégies Add&Remove et Transfer, ces probabilités ont été adaptées afin de rendre l'association entre le terme *mad* et la polarité négative plus forte. En effet, les probabilités associées à la polarité négative sont plus importantes que celles associées à la polarité positive.

3.4 Expérimentations

Afin de tester les performances de notre méthode, nous conduisons des expérimentations sur trois corpus différents décrits dans le tableau 3.7. MR est un ensemble de critiques anglophones par rapport à des films [85]. SemEval est un corpus de tweets anglophones sur des sujets variés, utilisé pour l'évaluation des systèmes participants à la campagne SemEval'2013 [122]. Critiques est un corpus de critiques francophones par rapport à des films, des livres et des hôtels [112].

Les paramètres de notre méthode hybride sont empiriquement fixés aux valeurs suivantes : pour la stratégie Add&Remove $\alpha_+ = \alpha_- = 0.002, \beta_+ = \beta_- = 0.3$ et pour la stratégie Transfer $\alpha_+ = \alpha_- = 0.3$.

Nous utilisons deux lexiques d'opinions : pour le Français, nous avons développé notre lexique en annotant manuellement 3927 termes dont 2697 positifs et 1230 négatifs [4]. Pour l'anglais, nous utilisons le lexique de B. Liu [5] que

4. Ressource accessible via la plateforme MediaMining à l'adresse : http ://mediamining.univ-lyon2.fr/

5. http ://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Corpus	Type	Langue	Taille	(pos./nég.)
MR	Critiques de films	Anglais	2000	(1000/1000)
SemEval	Tweets	Anglais	7666	(5431/2235)
Critiques	Critiques de films, livres et hôtels	Français	5034	(2658/2376)

TABLE 3.7 – Corpus utilisés pour l'analyse d'opinions.

nous avons enrichi avec des mots du langage informel (smileys, abréviations, etc.) collectés sur Urban Dictionary [6]. Le lexique résultant comporte 7720 termes dont 2475 positifs et 5245 négatifs. Comme prétraitement, nous utilisons uniquement la racinisation en se basant sur *Porter Stemmer* et la suppression des mots vides.

Prise en compte du biais. Afin d'examiner comment notre méthode prend en compte le problème de biais discuté dans la section 3.3.1, nous réalisons la même expérimentation que celle présentée dans le tableau 3.4 avec notre méthode hybride Add&Remove et nous comparons le résultat avec la méthode NB. Le tableau 3.8 présente les résultats obtenus avec la méthode hybride Add&Remove.

Terme (w)	Polarité *a priori*	Proba. donn. appr. ($\times 10^{-2}$)		Exemples d'erreurs de classement NB (c_r=classe réelle, c_p=classe prédite)		
		$p(w\|c_+)$	$p(w\|c_-)$	tweet	c_r	c_p
honesty	POS	0.23	0.12	There is an honesty that says we play the game Saturday..	c_+	c_-
critical	NEG	0.05	0.47	Need cough medicine and/or Lemsips. May be critical if you want the Gazette on Wednesday !..	c_-	c_-
fault	NEG	0.05	0.53	no matter how you slam Obama, you own our credit rating downgrade. It is ALL your fault..	c_-	c_-
mad	NEG	1.56	2.42	Tomorrow is KARWA- CHOTH Delhi is going MAD !!!!!!!..	c_-	c_-

TABLE 3.8 – Prise en compte du biais par notre méthode hybride Add&Remove. Ce tableau est lu en l'opposant au tableau 3.4.

Comme le montrent ces résultats, notre méthode hybride a permis de corriger le biais du à la sous-représentativité des données d'apprentissage. Tous les termes en situation de biais, présentés dans ce tableau, ont été ré-assignés à leurs classes *a priori* en corrigeant leurs probabilités. Cette correction s'est répercutée sur les résultats obtenus. En effet, la plupart des tweets qui ont été mal classés par la méthode NB ont été bien classés par la méthode hybride (cf. tableau 3.4). Le tableau 3.9 montre d'autres exemples de tweets qui ont été mal classés par la méthode NB et bien classés par notre méthode.

6. http ://www.urbandictionary.com/

Tweet	Classe réelle	NB	ADD& REMOVE
http ://t.co/EKOjBMU7 Siri vs. Google Voice Search-may the best robot helper win @iphonefirmware	c_+	c_-	c_+
Please, TS. Don't make me *dislike* Hyung Jun just like how I *dislike* SOME of you -,-	c_-	c_+	c_-
We 10th cuz the ACC is *garbage*. Big East has *tore* them up all year	c_-	c_+	c_-
Lots of students interested in UK study at the DEC fair in Kyiv today	c_+	c_0	c_+
@mkipley ND projected to finish 3rd in the Big East. Has a major size advantage of QU.	c_+	c_0	c_+
@CW_network *Bummed* We don't get to see Arrow until Saturday because of the Chicago Bulls game.	c_-	c_0	c_-
@laurensackett14 *nooooo* I am possibly going to be in dayton tomorrow night !!! *:(*	c_-	c_+	c_-
If I wasn't going to Owain's tomorrow , I would probably spend the whole night *crying* to myself *$:* ..	c_-	c_+	c_-
I get it now he told @clbrooks_48 tomorrow so she wouldn't come. Dang *nobody likes* Calle.	c_-	c_0	c_-
marth CASHEE retake tomorrow and i havent studied. still busy with this book report. *fml*	c_-	c_0	c_-
Oomf is gonna get *punched* in the throat tomorrow -.-	c_-	c_0	c_-
May the *odds* be ever in your favor. - The Hunger Games¨	c_+	c_0	c_+
Getting *sick* right before the UCLA football game on Saturday ¡¡¡	c_-	c_0	c_-

TABLE 3.9 – Une sélection de tweets qui sont mal classés par la méthode NB et bien classés par notre méthode.

Cependant, le problème n'est pas complètement résolu car certains tweets demeurent mal classés, comme le premier tweet avec le mot *honesty*. Cela est peut être du au fait que la correction de la probabilité pour le mot *honesty* n'a pas suffi à bien classer le tweet. Une des solutions pour aider le modèle à mieux appréhender ce type de situations serait d'augmenter de la valeur des paramètres responsables de l'adaptation de probabilités pour la classe positive α_+ et β_+.

Performances globales. Nous comparons notre méthode avec les méthodes NB et SVM. Pour ce dernier, deux configurations du noyau sont testées : linéaire (SVM-L) et polynomial de degré 2 (SVM-P).

Méthode	MR	SemEval	Critiques
NB	73.06	74.07	75.88
SVM-L	74.56	49.79	79.89
SVM-P	**84.64**	49.74	**86.67**
Add & Remove	80.57	**76.05**	86.58
Transfer	75.53	76.00	80.01

TABLE 3.10 – Résultats obtenus avec notre approche, NB et SVM (problème à deux classes).

Le tableau 3.10 représente les résultats obtenus sur les différents corpus, en terme de F-score avec validation croisée en 5 fois. Nous remarquons sur ce tableau que notre approche donne de meilleurs résultats par rapport à NB de base, et ce sur les trois corpus. Le gain en performance est variable et dépend du corpus utilisé (entre 2 et 11 points de F-score). Notre méthode demeure compétitive avec SVM.

Nous remarquons aussi que la stratégie Add&Remove donne de meilleurs résultats par rapport à la stratégie Transfer. Cela est du au fait que dans la première stratégie, seulement les termes déjà présents dans le corpus peuvent être transférés d'une classe à l'autre alors que dans la première stratégie, le nombre d'occurrences à rajouter est beaucoup plus important car ne dépend pas de la fréquence des termes dans le corpus.

Enfin, de manière plus générale, les résultats présentés dans le tableau 3.11 montrent que notre méthode est plus robuste au problème de sur-apprentissage. En effet, en plus d'améliorer le résultat global de classement, elle permet aussi de réduire l'écart entre les résultats obtenus sur les données d'apprentissage et ceux obtenus sur les données de test.

Méthode	Score	Données d'apprentissage	Données de test	Ecart (points de %)
NB	Fscore(c_+)	0.92	0.73	0.19
	Fscore(c_-)	0.58	0.49	0.09
	Fscore(c_0)	0.69	0.58	0.11
	Moyenne	**0.73**	**0.60**	**0.13**
Add&Remove	Fscore(c_+)	0.75	0.67	0.08
	Fscore(c_-)	0.63	0.59	0.04
	Fscore(c_0)	0.67	0.57	0.10
	Moyenne	**0.68**	**0.63**	**0.05**

TABLE 3.11 – Résultats obtenus avec les méthode NB et Add&Remove sur les données d'apprentissage et les données de test (corpus SemEval, problème à trois classes).

3.5 Discussion

Notre proposition pour améliorer les résultats de classement d'opinions avec la méthode NB a été fructueuse. En effet, les scores de classement sont nettement meilleurs par rapport à NB. Ces résultats nous ont permis de participer à la tâche 2 de la campagne d'évaluation SemEval-2013 [122] (classification supervisée d'opinions à partir des données Twitter) où notre système a été classé $6^{\text{ème}}/35$ avec les données Twitter et $9^{\text{ème}}/22$ avec les données SMS [7]. De plus, les scores obtenus avec notre méthode étaient supérieurs à la moyenne des scores de tous les participants (voir tableau 3.12). Ce classement est basé sur le score moyen du rappel obtenu sur les deux classes d'opinion positive et négative. Plus de détails sur les paramètres du modèle et les prétraitements des données sont données dans [20].

Corpus	Notre score	Moyenne des scores	Meilleur score	Plus faible score	Notre classement
Tweets	**62.55**	53.70	69.02	16.28	6/35
SMS	**53.63**	50.20	58.46	22.16	9/22

TABLE 3.12 – Résultats .

Cependant, malgré ces résultats, il est à noter que notre méthode ne permet pas de gérer l'opinion neutre, faute de lexique de termes neutres. La connaissance nécessaire à la détection du neutre provient seulement du corpus d'apprentissage. Il serait intéressant comme complément à ce travail de proposer des techniques et/ou des ressources qui permettraient de gérer les opinions neutres.

7. http ://www.cs.york.ac.uk/semeval-2013/

De manière générale, malgré les avancées réalisées dans ce domaine, l'analyse d'opinions demeure une tâche difficile. Contrairement aux thématiques qui sont généralement détectées par des mots clés, l'expression de l'opinion peut être implicite en utilisant le langage figuratif, comme dans les cas de la métaphore, du sarcasme et de l'ironie [94, 95, 110]. La détection de ces phénomènes requiert des outils plus élaborés que ceux classiquement déployés pour l'analyse d'opinions.

Le langage informel est un autre phénomène qui peut rendre plus compliquée la tâche d'analyse d'opinions. Dans les réseaux sociaux et les microblogs, comme l'exemple proéminent de Twitter, les internautes ont tendance à utiliser un langage informel composé d'acronymes, signes, argot, etc [122]. La présence de ces caractéristiques apporte une dimension supplémentaire à gérer lors de l'analyse d'opinions où les outils classiques peuvent s'avérer insuffisants. Durant ces dernières années, beaucoup de travaux se sont intéressés aux données Twitter en prenant en compte la spécificité du langage [48, 54, 84] et la structure [54].

Chapitre 4

Thématiques et Opinions : Modélisation Conjointe

Résumé. *Les thématiques et les opinions sont deux aspects du texte qui peuvent être traitées séparément ou conjointement. Dans ce chapitre, nous proposons un nouveau modèle probabiliste pour l'extraction conjointe des thématiques et des opinions à partir du texte. La problématique que nous proposons de résoudre est l'estimation des proportions d'opinions caractérisant les thématiques. Nous montrons à travers une évaluation adaptée la supériorité de notre modèle par rapport aux modèles de l'état de l'art en terme de prédiction des opinions au niveau de la thématique.*

Sommaire

4.1 Introduction

Après avoir traité la modélisation de thématiques et l'analyse d'opinions séparément dans les deux chapitres précédents, nous nous intéressons désormais à la jonction de ces deux aspects du texte. Comme cela a été dit dans l'introduction des chapitres 2 et 3, l'extraction de thématiques et l'analyse d'opinions ont des applications diverses, comme la recommandation, la veille sur le Web et la e-réputation. L'intérêt de ces techniques d'analyse est d'autant plus grand si ces deux aspects du texte sont traités conjointement, c'est à dire si l'extraction des thématiques et des opinions est réalisée simultanément. Cela permettrait de répondre à des questions telles que : "quel aspect du nouveau produit ne plaît pas aux clients ?", "quel est l'impact de la déclaration du Premier Ministre sur l'audience Twitter ?", "est-on toujours favorable à l'énergie nucléaire après l'incident nucléaire de Fukushima de 2011 ?".

La relation entre les thématiques et les opinions est indéniable. Les opinions sont généralement exprimées par rapport à des thématiques (cibles) et inversement, les thématiques peuvent être associées à des opinions. La cible d'une opinion peut prendre différentes formes [45] : explicite ("*les dimensions sont inacceptables*") ou implicite ("*l'appareil est trop grand*"). L'opinion peut concerner un objet manufacturé ("*un ordinateur de design élégant*"), un événement ("*belle soirée*"), une personne ("*elle n'était pas à la hauteur*"), etc. L'opinion peut concerner un objet en général ou un aspect particulier de cet objet ("*c'est un joli téléphone, seul bémol, son prix élevé*").

La problématique de modélisation conjointe des thématiques et des opinions a été abordée au début des années 2000 avec les premiers travaux dans l'analyse d'opinions [16, 45, 77]. On parlait de l'extraction de la cible d'opinion. Cette problématique a suscité un regain d'intérêt ces cinq dernières années avec l'arrivée des modèles probabilistes. Selon l'approche adoptée, le problème peut être vu soit comme une extension de la tâche d'extraction de thématiques afin d'inclure les opinions, soit comme une extension de la tâche d'analyse d'opinions afin de détecter aussi les thématiques, appelées les cibles d'opinion. Dans les deux cas, l'objectif est le même : extraire les paires thématiques-opinions. Afin de décrire formellement la problématique, nous proposons la définition suivante :

Définition 4 (extraction conjointe thématiques-opinions) *La problématique d'extraction conjointe des thématiques et des opinions consiste à trouver l'ensemble des paires* $\mathbb{P} = \{(z, o)\}$ *à partir d'une collection de documents, où* z *représente une thématique et* o *caractérise l'opinion relative à* z*, par exemple par une distribution sur les polarités de l'opinion.*

Dans ce chapitre, nous proposons un nouveau modèle probabiliste TS (*Topic-Sentiment model*) pour l'extraction conjointe des thématiques et des opinions.

Notre modèle permet d'estimer les proportions d'opinions pour chacune des thématiques extraites. A travers des expérimentations et des évaluations adaptées, nous montrerons comment cette caractéristique permet de mieux prédire l'opinion au niveau de la thématique. Ces travaux ont été publiés dans [21].

Dans les deux sections qui suivent, nous présentons l'état de l'art du domaine de l'extraction conjointe des thématiques et des opinions ainsi que les méthodologies d'évaluation. Dans la section 4.4, nous présentons notre contribution. Dans la section 4.5, nous décrivons les expérimentations et les résultats. Enfin, dans la section 4.6, nous discutons les résultats et nous donnons quelques perspectives.

4.2 Etat de l'art

Quelques travaux de l'analyse d'opinions, notamment des travaux basés sur les règles linguistiques, ont été étendus pour l'extraction des cibles d'opinions [45, 89, 104]. Plus récemment, nous avons assisté à l'apparition d'un courant de travaux basés sur les modèles de thématiques probabilistes pour l'extraction conjointe des opinions et des des cibles [37, 49, 52, 59, 60, 63, 64, 70, 115]. Les thématiques sont dans ce cas considérées comme des cibles. Nous proposons de catégoriser l'ensemble de ces travaux en deux approches : approche *post hoc* et approche conjointe. Dans l'approche *post hoc*, on procède à l'extraction des thématiques et des opinions en deux temps, de manière séquentielle, avant de les mettre en correspondance. Dans l'approche conjointe, on effectue l'extraction des thématiques et des opinions de manière simultanée tout en prenant en compte les associations entre elles (figure 4.2.1).

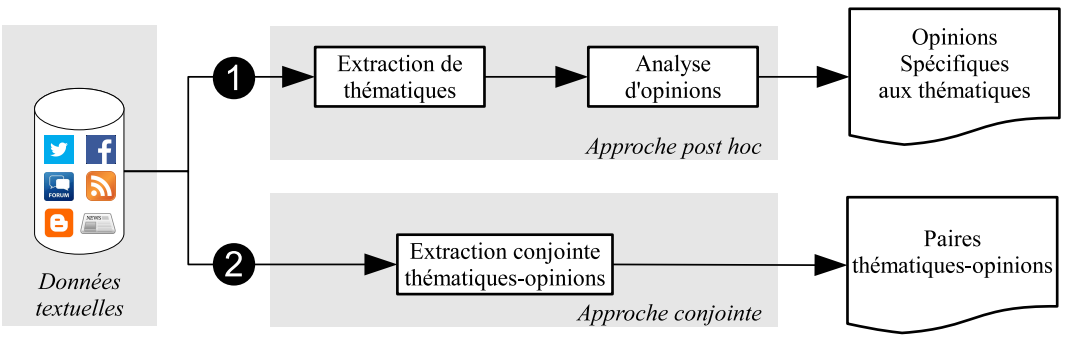


FIGURE 4.2.1 – Catégorisation en deux approches des méthodes d'extraction conjointe thématiques-opinions.

4.2.1 Approche post hoc

Les méthodes *post hoc* sont généralement basées sur l'utilisation des règles linguistiques (cf. section 3.2.2). Les méthodes déjà développées pour l'analyse d'opinions sont généralement reprises et enrichies afin de détecter, en plus des opinions, les cibles de celles-ci. Cette approche est connue sous l'appellation *feature/aspect-based sentiment analysis*. Ainsi, l'extraction d'opinions est précédée d'une phase d'extraction de cibles (*"features"*), qu'on peut considérer comme des thématiques explicites, dans une démarche *post hoc*. Conformément à la littérature, nous utilisons ici le terme "cible" pour désigner une thématique faisant l'objet d'une opinion.

Hu et Liu [45] sont parmi les premiers à s'être intéressés à cette problématique. Ils ont proposé d'agréger les opinions en regroupant celles qui portent sur la même cible. Pour cela, ils se sont basés sur un corpus de critiques de produits où chaque produit est décrit par un ensemble de caractéristiques. L'approche générale se déroule comme suit :

1. Les caractéristiques des produits sont extraites en utilisant un analyseur syntaxique pour la détection des groupes nominaux fréquents, ceux-ci étant considérés comme les caractéristiques (cibles). Par exemple, pour un appareil photo numérique, les cibles sont : "*picture quality*", "*size*", "*battery life*", etc.
2. Toutes les expressions d'opinions proches de ces cibles sont ensuite identifiées. Une expression d'opinion est composée d'une cible et d'un adjectif. Ensuite, la polarité de chaque expression est déterminée en utilisant une méthode basée sur WordNet, similaire à [50].
3. Enfin, les résultats sont agrégés et présentés pour chaque cible.

Popescu et Etzioni [89] proposent OPINE : un système pour l'extraction d'opinions et de cibles. Cette approche est très similaire à la précédente [45] mais utilise la ressource WordNet afin d'améliorer la détection des cibles. Somprasertsri et Lalitrojwong [104] utilisent les règles syntaxiques et sémantiques en plus de relations de dépendance. Ils se sont également servis d'ontologies pour l'extraction des opinions et des cibles avec un algorithme d'apprentissage supervisé similaire à *Naive Bayes*.

La limite principale de ce type de méthodes réside dans la définition de la cible de l'opinion. Ces méthodes considèrent les thématiques comme étant des caractéristiques de produits, souvent exprimées avec des phrases nominales ou détectées sur la base d'une liste ou d'une ontologie. Il est souvent impossible de généraliser cette définition pour inclure des thématiques au sens large du terme. Chacune de ces approches est destinée à une application précise. En outre, ces méthodes nécessitent des ressources linguistiques (analyseurs syntaxiques, ontologies, etc.) ainsi que des règles de classement coûteuses à créer et très dépendantes du domaine [68].

4.2.2 Approche conjointe

L'approche conjointe repose sur l'hypothèse qu'il existe des associations et des liens entre les thématiques et les opinions qui font qu'il devient plus judicieux que ces deux aspects soient extraits ensemble. Cette approche a été essentiellement étudiée ces dernières années avec l'apparition des modèle de thématiques probabilistes comme PLSA et LDA [9, 44].

Comme nous l'avons déjà fait remarquer dans le chapitre 2, les modèles de thématiques probabilistes se basent sur la co-occurrence des termes afin d'extraire les thématiques (ensembles de termes co-occurrents). Pour la modélisation conjointe, ce principe est étendu. Afin d'extraire les conjonctions thématiques-opinions, les modèles conjoints se basent sur la co-occurrence entre les termes exprimant des thématiques et les termes exprimant des opinions.

Pour illustrer ce principe, nous nous appuyons sur les travaux de Lin et He [63, 64] datant de 2009. Ces travaux ont été précédés de quelques tentatives de modéliser conjointement les thématiques et les opinions, notamment [70]. Dans [63, 64], les auteurs proposent le modèle JST (*Joint Sentiment-Topic model*). Ce modèle peut être vu comme une extension du modèle LDA avec un niveau hiérarchique plus général que la thématique. Ce niveau regroupe les thématiques traitées sous la même polarité d'opinion. Une nouvelle variable aléatoire latente l est dédiée à l'extraction des opinions relatives aux thématiques. Comme le montre la figure 4.2.2, la variable des thématiques z dépend de la variable de l'opinion l. Cela est équivalent à un modèle LDA avec deux niveaux hiérarchiques où le premier niveau correspond à la polarité de l'opinion et le deuxième niveau correspond à la thématique. Plus de détails sur le fonctionnement des modèles probabilistes ont été donnés dans le chapitre 2.

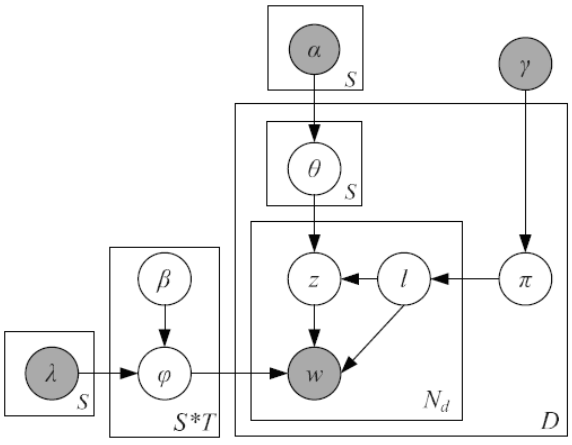


FIGURE 4.2.2 – Modèle graphique de JST [63].

Le modèle JST utilise le même principe que LDA. Les distributions de probabilités caractérisant les thématiques sont de la même nature (multino-

miales conditionnées par des lois de Dirichlet). Afin d'estimer les opinions relatives aux documents, JST utilise une distribution supplémentaire (π) qui est également de nature multinomiale car l'opinion est considérée comme une variable catégorielle. La variable de Dirichlet γ est utilisée afin d'exprimer un *a priori* pour la répartition des documents sur les polarités d'opinion (de la même manière que les variables α et β).

Le processus génératif du modèle JST correspondant au graphe de la figure 4.2.2 se déroule comme suit :

1. Pour chaque document d, se fixer une distribution sur les opinions $\pi_d \sim$ Dirichlet(γ)
2. Pour chaque opinion l et une thématique z, se fixer une distribution sur le vocabulaire $\varphi_{l,z} \sim$ Dirichlet(β)
3. Pour chaque opinion l et un document d, se fixer une distribution sur les thématiques $\theta_{d,l} \sim$ Dirichlet(α)
4. Pour chaque terme w_i dans le document d
   (a) Tirer une opinion $l_i \sim \pi_d$
   (b) Tirer une thématique $z_i \sim \theta_{d,l_i}$
   (c) Tirer un terme w_i de la distribution sur le vocaulaire φ_{l_i,z_i} spécifique à la thématique z_i et à l'opinion l_i

Dans JST, l'extraction d'une thématique est conditionnée par la connaissance de la polarité d'opinion. L'association d'une thématique avec une opinion est basée sur la co-occurrence des termes thématiques avec les termes porteurs d'opinions. Afin de distinguer les termes porteurs d'opinions, le modèle se base sur une liste noyau (*seed list*) composée de termes polarisés (comme dans les méthodes à base de similarité entre les termes, cf. section 3.2.3), d'où le caractère "faiblement supervisé" de ce modèle. Les thématiques sont construites en se basant sur la co-occurrence des termes polarisés avec les termes non polarisés.

Dans [64], les auteurs de JST ont testé *Reverse*-JST, un modèle similaire à JST où l'ordre des nœuds z et l est inversé (cf. figure 4.2.3). Dans *Reverse*-JST, le premier niveau de hiérarchie correspond aux thématiques et le deuxième niveau correspond aux opinions. Ainsi, une thématique est caractérisée par une distribution de probabilités sur les opinions, spécifique à chaque document.

Le modèle JST a inspiré la proposition de plusieurs autres modèles autour de la même problématique. Nous présentons ces modèles de manière chronologique en commençant par les plus récents. A la fin de cette section, nous présentons une discussion récapitulative qui permet de mettre en valeur les similarités et les différences entre ces modèles.

Sentiment-LDA [60] est un modèle similaire à JST mais il prend en compte seulement deux polarités d'opinions (positive et négative). La distribution des

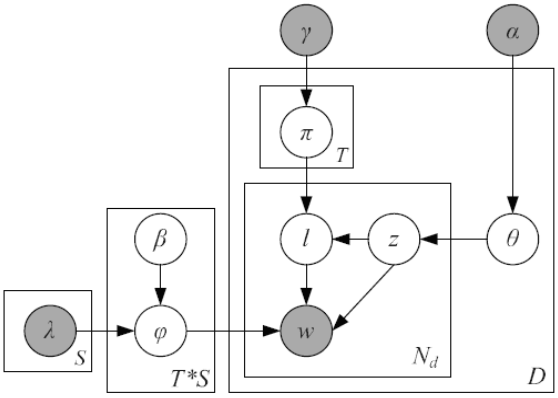


FIGURE 4.2.3 – Modèle graphique de *Reverse*-JST [64].

thématiques sur les opinions est caractérisée par une loi de Bernoulli contrairement à JST où cette distribution suit une loi multinomiale. Dans *Dependency-Sentiment-LDA* [60], la relation entre des mots de la même phrase, perdue avec l'utilisation du modèle vectoriel, a été rétablie en partie. Deux mots d'opinion de la même phrase peuvent ainsi exprimer des opinions "similaires" ou "opposées". Le changement d'opinion au sein de la même phrase est modélisé par une chaîne de Markov.

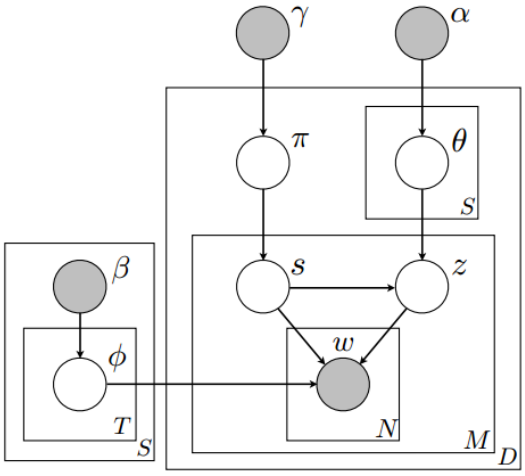


FIGURE 4.2.4 – Modèle graphique de ASUM [49].

ASUM (*Aspect and Sentiment Unification Model*) [49] est un autre modèle similaire à JST basé sur une hypothèse suivante : les mots de la même phrase sont forcément liés à la même thématique et expriment la même opinion. Ainsi, dans le processus génératif de ASUM, les mots de la même phrase sont tous tirés avec la même loi multinomiale, i.e. la même thématique et la même pola-

rité d'opinion. Un document est découpé en M phrases dont chacune comporte un nombre variable de mots N (cf. figure 4.2.4).

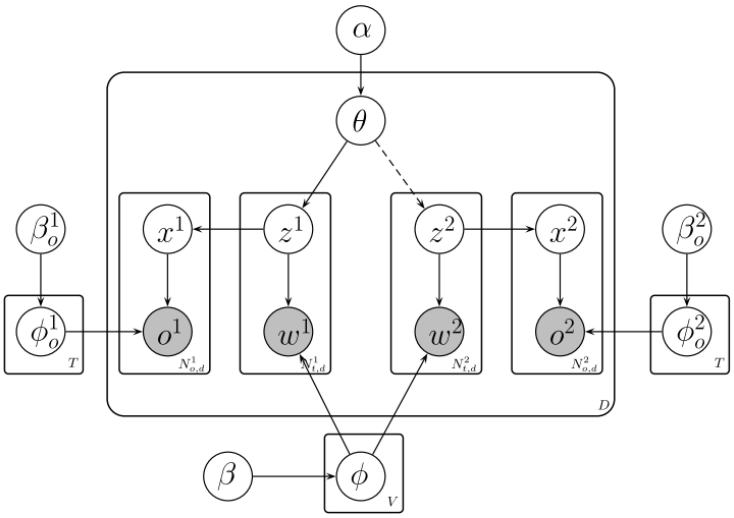


FIGURE 4.2.5 – Modèle graphique de CPT [31].

CPT (*Cross-perspective Topic Model*) est un modèle légèrement différent par rapport aux modèles cités ci-dessus [31]. Comme le montre la figure 4.2.5, CPT combine deux sous-modèles pour l'extraction de thématiques sous deux perspectives d'opinions (positive et négative). L'arc en pointillés sur la figure est utilisé pour montrer qu'un document ne peut venir que de l'une de ces deux polarités d'opinion. Ce modèle peut donc être considéré comme supervisé car l'information sur la polarité du document est explicitement utilisé par le modèle. Cependant, la collection des documents est utilisée en entier pour l'extraction de thématiques. Celles-ci ont différentes distributions sur le vocabulaire selon la polarité de l'opinion.

STDP (*Sentiment Topic with Decomposed Prior*) [59] est un autre modèle très similaire à JST. L'association d'une opinion à un mot est décomposée en deux étapes : (1) en se basant sur sa catégorie grammaticale (POS), déterminer si le mot est un mot d'opinion ou pas. (2) Si oui, tirer une opinion pour ce mot.

HASM (*Hierarchical Aspect-Sentiment Model*) [52] est un modèle hiérarchique qui associe les thématiques et les opinions (cf. figure 4.2.6). HASM adopte la même démarche que *Reverse*-JST (une même thématique est extraite pour les deux polarités d'opinion) et la même granularité de données que ASUM (découpage des documents en phrases). Le résultat de HASM se présente comme un arbre binaire où chaque nœud parent est une thématique générale des deux nœuds enfants. En plus, chaque nœud est aussi caractérisé par deux distributions sur le vocabulaire, une pour chaque polarité d'opinion. En revanche, le modèle ne permet pas de caractériser les proportions d'opinions

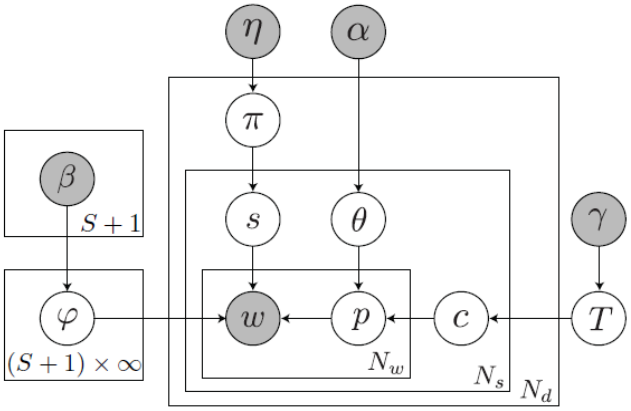


FIGURE 4.2.6 – Modèle graphique de HASM [52].

relatives aux thématiques. Celles-ci sont seulement calculées au niveau des phrases.

Gottipati et al. [37] ont proposé d'étendre le modèle LDA avec une variable d'opinion et quatre autres variables afin de capturer la spécificité de l'opinion : terme exprimant une opinion relative à la thématique, terme exprimant une opinion générale, terme relatif à la thématique, terme vide.

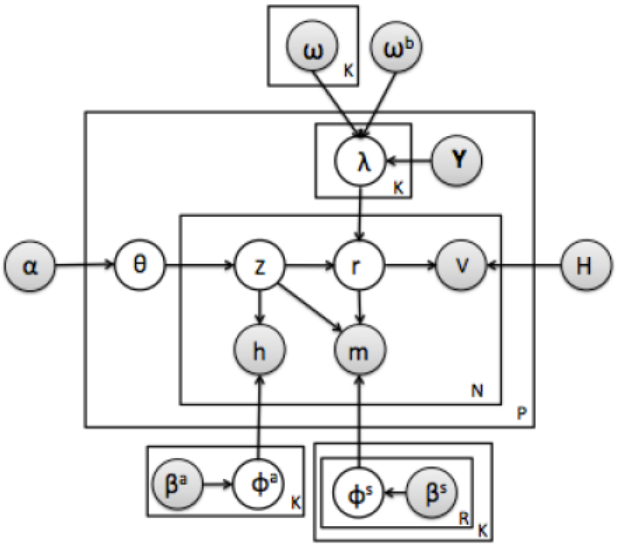


FIGURE 4.2.7 – Modèle graphique de SATM [115].

Plus récemment, Wang et Ester ont proposé SATM (*Sentiment-Aligned Topic Model*) [115] pour la prédiction des opinions relatives aux caractéristiques de produits. Le modèle s'appuie sur deux types de connaissances *a priori* : un lexique d'opinion et une distribution sur les opinions, spécifique à chaque

produit. Pour chaque paire (produit, caractéristique), le modèle fournit une distribution sur les polarités de l'opinion (cf. figure 4.2.7). Cette distribution est paramétrée par la connaissance *a priori*. Contrairement aux autres modèles, SATM prend aussi comme entrée la structure des données (produits, caractéristiques).

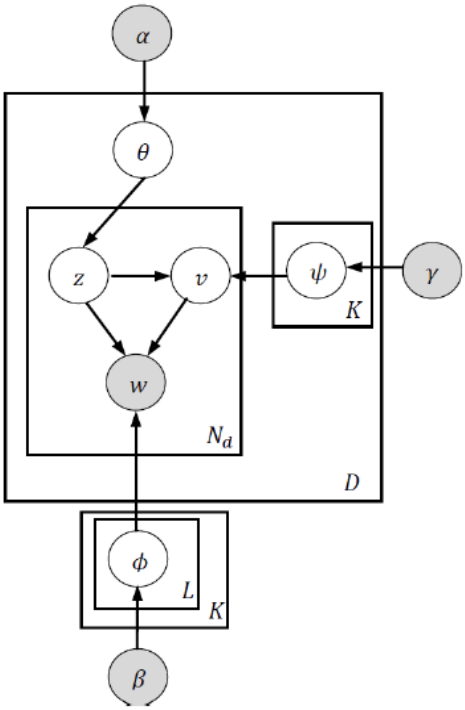


FIGURE 4.2.8 – Modèle graphique de JTV [109].

JTV (*Joint Topic Viewpoint model*) [108, 109] est un modèle très similaire à *Reverse*-JST où les thématiques sont caractérisées par des distributions sur les polarités d'opinion. Contrairement à *Reverse*-JST, il n'y a aucune connaissance *a priori* qui guide la détection des opinions. Par conséquent, le modèle n'est pas initialement destiné à extraire des polarités d'opinions mais des points de vue contrastés par rapport à une thématique, comme le support et la détraction.

4.2.3 Discussion

Les modèles présentés dans cette section comportent de nombreuses similarités. Afin de discuter plus en détails l'ensemble de ces modèles, nous les représentons dans le tableau récapitulatif 4.1. Nous proposons quelques critères qui permettent de comparer ces modèles et de mettre en valeur les différences et les spécificités de chacun par rapport aux autres. Ces critères sont les suivants :

– **Premier niveau hiérarchique :** certains modèles permettent l'extraction des thématiques sous différentes polarités d'opinions, c'est-à-

Modèle et référence	Année	Basé sur	Premier niveau hiérarchique	Portée de l'opinion niveau thém.	Supervisé (niveau document)	Lexique	Unité thématique	opinion niveau document
TSM [70]	2007	PLSA	opinion	document	non	oui	terme	non
JST [63, 64]	2009	LDA	opinion	document	non	oui	terme	oui
Sentiment-LDA [60]	2010	LDA	thém.	document	non	oui	terme	non
ASUM [49]	2011	LDA	opinion	document	non	oui	phrase	oui
Reverse-JST [64]	2012	LDA	thém.	document	non	oui	terme	non
CPT [31]	2012	LDA	thém.	corpus	oui	non	terme	non
STDP [59]	2013	LDA	opinion	document	non	oui	terme	oui
HASM [52]	2013	LDA	thém.	document	non	oui	phrase	oui
Gottipati et al. [37]	2013	LDA	opinion	document	non	oui	terme	oui
SATM [115]	2014	LDA	thém.	produit	oui	oui	terme	non
JTV [108, 109]	2014	LDA	thém.	document	non	non	teme	non

TABLE 4.1 – Modèles probabilistes pour l'extraction conjointes des thématiques et des opinions.

dire chaque thématique est affectée à une opinion de manière définitive. D'autres modèles commencent d'abord par l'extraction des thématiques puis la caractérisation de ces thématiques par des distributions sur les polarités d'opinion. Ce critère concerne le premier niveau hiérarchique du modèle (thématique ou opinion).

– **Portée des opinions relatives aux thématiques :** les opinions relatives aux thématiques, qu'elles soient exprimées de manière définitive ou par des distributions de probabilité, peuvent être estimées globalement (pour tout le corpus) ou spécifiquement (pour chaque document).
– **Supervision :** certains modèles conjoints sont supervisés, c'est-à-dire qu'ils utilisent explicitement l'annotation des documents avec les opinions pour apprendre.

- **Lexique :** une partie des modèles présentés dans cette section s'appuient sur un lexique d'opinion qui sert à guider la détection et le classement des opinions relatives aux thématiques et/ou aux documents.
- **Unité thématique :** nous appelons "unité thématique" le plus grand segment d'un document qui exprime la même thématique. Certains modèles se basent sur les termes (mots, n-grammes), d'autres se basent sur les phrases où tous les mots d'une phrase sont liés à la même thématique.
- **Prédiction de l'opinion au niveau des documents :** certains modèles fournissent en plus de l'opinion au niveau des thématiques, l'opinion au niveau de chaque document, même si ce n'est pas l'objectif premier de ce type de modèles.

Parmi tous ces critères, soulignons l'importance du "premier niveau hiérarchique" (thématique ou opinion). En effet, la forme du résultat fourni par le modèle n'est plus la même dans les deux cas. Selon ce critère, nous distinguons deux types de modélisation : "des thématiques sous des opinions" et "des opinions pour des thématiques".

Des thématiques sous des opinions. Ces modèles procèdent par l'extraction des thématiques sous des polarités d'opinions. Chaque thématique extraite est affectée à une opinion de manière définitive. Dans le modèle graphique, la variable des opinions dépend de la variable des thématiques, c'est-à-dire que la génération d'une opinion est conditionnée par la connaissance de la thématique. Ce type de modèles présente deux limites majeures que nous résumons comme suit :

1. Dans les données du Web, une thématique est généralement discutée sous différentes perspectives avec différentes proportions. Or, dans ce type de modélisation, ce phénomène n'est pas pris en compte. Dans [63], les auteurs ont réussi à extraire une thématique relative au film *Titanic* (James Cameron, 1997) avec le modèle JST à partir d'un corpus de critiques de films. Cette thématique est affectée à une polarité positive. Cela suppose qu'elle n'a été traitée que de manière positive, ce qui n'est pas le cas car une partie des critiques était bel et bien négative.
2. Aucune correspondance n'existe entre les thématiques sémantiquement similaires mais qui apparaissent sous différentes polarités d'opinion. Seul un travail manuel de post-traitement permet de rapprocher ces thématiques entre elles.

Des opinions pour des thématiques. Dans cette catégorie de modèles, les opinions dépendent des thématiques. Chaque thématique extraite est caractérisée par une distribution de probabilités sur les polarités d'opinion. Cette approche, plus réaliste, permet de remédier aux problèmes cités précédemment.

En revanche, tous les modèles de cette catégorie partagent un inconvénient qui concerne la spécificité de ces distributions. En effet, comme le montre le tableau 4.1, la distribution d'une thématique sur les polarités d'opinion est spécifique à chaque document. Ces modèles ont été conçus de manière à optimiser la prédiction de l'opinion pour les documents. Mais même si ce n'est pas l'objectif premier de ce type de modèles, il devient plus intéressant de généraliser ces distributions afin qu'elles portent sur tout le corpus globalement et non sur chaque document pris séparément.

4.3 Evaluation

Les modèles conjoints pour l'extraction des thématiques et des opinions ont été évalués de différentes manières qui permettent de prendre en compte la dimension de l'opinion.

Prédiction de l'opinion au niveau du document. La plupart des travaux cités dans la section 4.2.2 ont été évalués sur la base de la prédiction de l'opinion au niveau du document ou de la phrase. Ainsi, le modèle est utilisé pour prédire l'opinion pour un ensemble de documents qui ne sont pas utilisés en apprentissage. Cette mesure, certes intuitive et facilement interprétable, a deux limites principales :

- Tous les modèles ne permettent pas la prédiction de l'opinion au niveau du document, i.e., le résultat doit être post-traité afin d'estimer cette information
- La prédiction de l'opinion n'est pas la vocation initiale de ce type de méthodes. En effet, cette mesure ne prend pas en compte l'association entre les opinions et les thématiques. Un bon modèle pour prédire l'opinion au niveau du document n'est pas forcément un bon modèle pour extraire les associations entre les opinions et les thématiques.

Autres méthodes d'évaluation. Mise à part la mesure de perplexité, utilisée dans plusieurs travaux de modélisation conjointe [31, 108, 109], et dont les limites ont déjà été discutées dans [13], d'autres techniques sont utilisées pour évaluer ce type de modèles. Les travaux dans [52, 108, 109] utilisent des évaluations à base de dissimilarité ou de divergence pour mesurer l'hétérogénéité entre les thématiques mais cela ne prend pas en compte les associations thématiques-opinions. Dans [31], les auteurs se sont appuyés sur des mesures issues de la recherche d'information (capacité du modèle à fournir des résultats de recherche pertinents en réponse à des une requête). Enfin, une mesure spécifique à la prédiction de l'opinion relative aux caractéristiques d'un produit a été utilisée dans [115] mais celle-ci n'est valable que pour les thématiques de type produit/caractéristique.

4.4 Contribution : le modèle TS (*Topic-Sentiment model*)

Comme discuté dans la section 4.2.3, les modèles probabilistes proposés pour l'extraction conjointe des thématiques et des opinions partagent un inconvénient majeur : aucun de ces modèles ne permet d'estimer les proportions d'opinions pour les thématiques extraites. Certains modèles comme ceux présentés dans [52, 60, 64, 108, 115] permettent d'extraire cette information mais de manière spécifique, c'est-à-dire pour chaque document. Or, la connaissance des proportions d'opinions relatives aux thématiques est potentiellement importante pour avoir une vue d'ensemble des associations thématiques-opinions dans un corpus de documents.

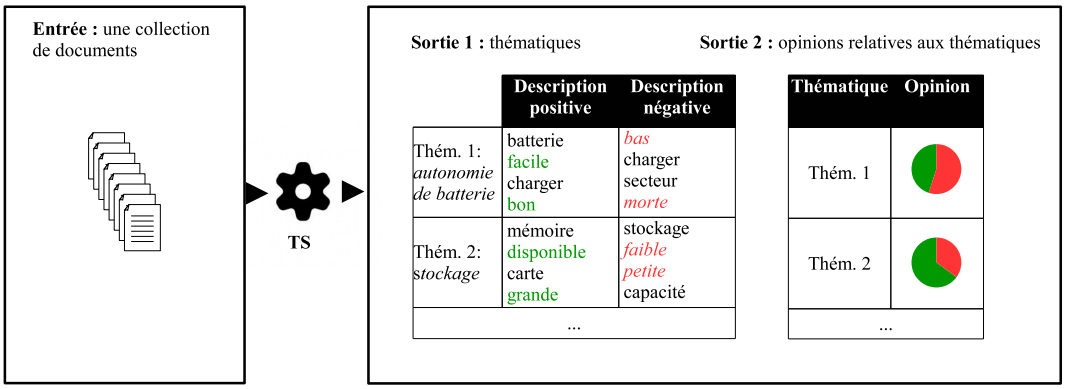


FIGURE 4.4.1 – Modélisation des thématiques et des opinions avec le modèle TS.

Pour traiter ce problème, nous proposons un nouveau modèle probabiliste, appelé TS (*Topic-Sentiment model*), qui permet d'extraire les thématiques et de les caractériser par les proportions d'opinions qui y sont associées comme l'illustre la figure 4.4.1. Pour ce faire, notre modèle est basé sur l'utilisation d'une distribution multinomiale spécifique à chaque thématique qui caractérise son association avec les polarités d'opinion. Afin de fournir une vue globale, ces distributions sont estimées à partir de tous les documents du corpus à la fois.

Nous avons opté pour les modèles probabilistes pour plusieurs raisons. D'abord pour leur performance par rapport aux autres modèles concurrents, comme par exemple les modèles à base de factorisation de matrices. En effet, les travaux d'évaluation que nous avons conduits (décrits dans le chapitre 2) ont montré la bonne performance du modèle LDA par rapport à deux méthodes : NMF (basée sur la factorisation de matrices) et FCM (basée sur le groupement par partitionnement). Qui plus est, les modèles probabilistes ont l'avantage d'être aisément extensibles avec de nouvelles variables afin de capturer

d'autres aspects du texte. L'extension du modèle LDA a déjà de bons résultats dans plusieurs domaines et notamment dans le domaine de l'analyse d'opinions [49, 59, 60, 63, 115]. En outre, et contrairement aux approches post hoc, la modélisation conjointe des thématiques et des opinions permet de prendre en compte les interactions entre elles. Ces deux aspects du texte s'influencent mutuellement à cause de la co-occurrence des mots thématiques avec les mots porteurs d'opinions. Une modélisation conjointe permet de mieux appréhender ces interactions et ainsi de mieux modéliser les associations thématiques-opinions dans leur ensemble.

4.4.1 Modèle graphique et processus génératif

Le modèle graphique de TS est représenté par la figure 4.4.2. Toutes les notations sont définies dans le tableau 1. Les variables w, z et s représentent respectivement les termes (mots, n-grammes, etc.), les thématiques et les opinions. La disposition des variables z et s dans cet ordre (la variable s dépend de la variable z) rend le modèle capable d'estimer des opinions pour des thématiques, c'est-à-dire que le premier niveau hiérarchique correspond à la thématique. La variable π représente les distributions multinomiales qui caractérisent les proportions d'opinions relatives aux thématiques. Cette variable se trouve en dehors du cadre D car elle n'est pas répétée pour chaque document. En d'autres termes, cette configuration permet de rendre ces distributions globales.

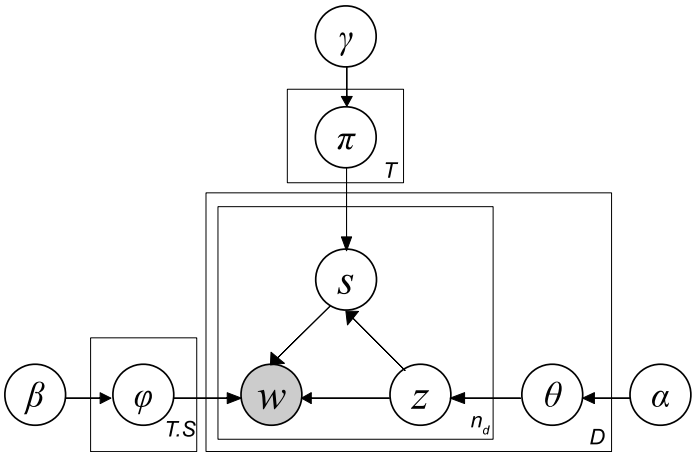


FIGURE 4.4.2 – Modèle graphique de TS.

TS, comme tout autre modèle probabiliste génératif, peut être considéré comme un processus génératif de documents. Le processus génératif correspondant au modèle graphique de la figure 4.4.2 se déroule comme suit :

1. Tirer $T \times S$ multinomiales $\varphi_{z,s} \sim \text{Dirichlet}(\beta)$
2. Tirer T multinomiales $\pi_z \sim \text{Dirichlet}(\gamma)$

3. Pour chaque document d, tirer une multinomiale $\theta_d \sim \text{Dirichlet}(\alpha)$, puis pour chaque terme w_i dans d :
   (a) Tirer une thématique $z_i \sim \theta_d$
   (b) Tirer une opinion $s_i \sim \pi_{z_i}$
   (c) Tirer un terme $w_i \sim \varphi_{z_i,s_i}$

Le déploiement du modèle TS pour l'extraction conjointe des thématiques et des opinions se fait en inversant ce processus génératif, c'est-à-dire, en cherchant à estimer les paramètres du modèle (les distributions multinomiales φ, θ et ψ) à partir des variables observées (w). Cette opération peut être réalisée en utilisant une méthode d'estimation des paramètres (inférence) parmi celles évoquées dans la section 2.3.1.

4.4.2 Inférence

Afin de réaliser l'inférence pour le modèle TS, nous utilisons la méthode d'échantillonnage de Gibbs (*Gibbs Sampling*) [39, 49, 59, 60, 63, 64]. La méthode d'échantillonnage de Gibbs est devenue très populaire avec l'apparition de modèles probabilistes de plus en plus complexes. En effet, quand l'inférence exacte est impossible ou est très compliquée, la méthode d'échantillonnage de Gibbs constitue une bonne alternative de manière générale.

La procédure d'échantillonnage de Gibbs est assimilée à une chaîne de Markov où l'état du modèle à un instant donné dépend uniquement de son état à l'instant précédent. Nous entendons par l'état du modèle l'association de tous les termes du corpus à des thématiques et aux opinions, c'est-à-dire les valeurs des paramètres φ, θ et ψ. Ainsi, un état initial est généralement défini de manière aléatoire puis, à chaque itération, chaque terme est ré-associé à une thématique et une opinion en fonction de toutes les autres associations. Cette opération est réalisée en se basant sur les valeurs des paramètres φ, θ et ψ calculées précédemment.

Le calcul de ces paramètres est une opération à deux étapes : nous avons d'abord besoin de calculer la distribution jointe des termes, des thématiques et des opinions : $p(\mathbf{w}, \mathbf{s}, \mathbf{z}|\alpha, \beta, \gamma)$. Ensuite, à partir de cette distribution nous déduisons la distribution *a posteriori* qui correspond à la probabilité de tirer une thématique et une opinion pour un terme donné quand les paramètres du modèle sont connus. Avant de détailler ces deux étapes, nous commençons par donner des généralités et des définitions indispensables à la réalisation de l'inférence.

Généralités sur la distribution de Dirichlet. La distribution de Dirichlet de paramètre a, notée Dirichlet(a), est une famille de distributions continues, paramétrée par le vecteur de réels positifs a de dimension $K \geq 2$. Elle est

la conjuguée de la loi multinomiale. Pour une distribution multinomiale $\mathbf{p}$ de dimension K, la loi de Dirichlet est définie comme suit :

$$\text{Dirichlet}(\mathbf{p}|a) = \frac{1}{\Delta(a)} \prod_{i=1}^{K} p_i^{a_i - 1} \tag{4.1}$$

où Δ représente la fonction Delta définie par la formule suivante :

$$\Delta(a) = \frac{\prod_{i=1}^{K} \Gamma(a_i)}{\Gamma(\sum_{i=1}^{K} a_i)} \tag{4.2}$$

où Γ représente la fonction Gamma. Cette fonction peut être vue comme la généralisation de la factorielle pour l'ensemble des réels et des complexes. Cette fonction a la propriété de récursivité suivante (pour $x \in \mathbb{C}$ de partie réelle strictement positive) :

$$\Gamma(x+1) = x \cdot \Gamma(x) \tag{4.3}$$

Cette propriété est importante car elle nous permet de simplifier le calcul de la distribution conditionnelle par la suite.

L'utilisation de la fonction Δ nous permet de calculer les probabilités nécessaires pour l'échantillonnage de Gibbs grâce à sa deuxième définition donnée ci-dessous :

$$\Delta(a) = \int_{\sum_i x_i = 1} \prod_{i=1}^{K} x_i^{a_i - 1} \, \mathrm{d}^K \mathbf{x} \tag{4.4}$$

Distribution jointe. En utilisant la règle d'indépendance conditionnelle et le modèle de la figure 4.4.2, la probabilité jointe d'observer un terme, une thématique et une opinion peut être décomposée comme suit :

$$p(\mathbf{w}, \mathbf{s}, \mathbf{z}|\alpha, \beta, \gamma) = p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \beta) \cdot p(\mathbf{s}|\mathbf{z}, \gamma) \cdot p(\mathbf{z}|\alpha). \tag{4.5}$$

Les facteurs de la distribution jointe peuvent être traités séparément. Pour obtenir le premier facteur, cherchons d'abord la distribution $p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \varphi)$. Celle-ci peut être directement calculée à partir de φ en multipliant les contributions provenant des termes et des opinions :

$$p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \varphi) = \prod_{j=1}^{T} \prod_{k=1}^{S} \prod_{i=1}^{V} \varphi_{j,k,i}^{n_{i,j,k}} \tag{4.6}$$

Les indices i, j, k sont utilisés pour indexer les termes, les thématiques et les opinions respectivement.

Le premier facteur de l'équation 4.5 est obtenu en intégrant par rapport à φ :

$$p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \beta) = \int p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \varphi) \cdot p(\varphi|\beta) \, \mathrm{d}\varphi \tag{4.7}$$

En utilisant la loi de probabilité de Dirichlet (équation 4.1) et l'équation 4.6, on obtient :

$$\begin{aligned} p(\mathbf{w}|\mathbf{s},\mathbf{z},\beta) &= \int \prod_{j=1}^{T}\prod_{k=1}^{S}\left(\prod_{i=1}^{V}\varphi_{j,k,i}^{n_{i,j,k}}\cdot\frac{1}{\Delta(\beta)}\prod_{i=1}^{V}\varphi_{j,k,i}^{\beta_i-1}\right)\mathrm{d}\varphi_{j,k} \\ &= \prod_{j=1}^{T}\prod_{k=1}^{S}\left(\frac{1}{\Delta(\beta)}\cdot\int\prod_{i=1}^{V}\varphi_{j,k,i}^{n_{i,j,k}+\beta_i-1}\mathrm{d}\varphi_{j,k}\right) \end{aligned} \tag{4.8}$$

En réécrivant l'intégrale de Dirichlet en fonction de Δ (équation 4.4) :

$$p(\mathbf{w}|\mathbf{s},\mathbf{z},\beta) = \prod_{j=1}^{T}\prod_{k=1}^{S}\frac{1}{\Delta(\beta)}\cdot\Delta(n_{i,j,k}+\beta) \tag{4.9}$$

En remplaçant maintenant la fonction Δ par sa première définition utilisant la fonction Γ (équation 4.2), nous retrouvons la formule finale permettant de calculer le premier facteur de la probabilité jointe :

$$p(\mathbf{w}|\mathbf{s},\mathbf{z},\beta) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{T\cdot S}\prod_{j}\prod_{k}\frac{\prod_i\Gamma(n_{i,j,k}+\beta)}{\Gamma(n_{j,k}+V\beta)}, \tag{4.10}$$

Les écritures $\Gamma(V\beta)$ et $\Gamma(\beta)^V$ sont équivalentes car l'hyperparamètre β est symétrique.

Les facteurs restants de l'équation 4.5 sont obtenus de manière identique en intégrant par rapport à π et θ respectivement

$$p(\mathbf{s}|\mathbf{z},\gamma) = \left(\frac{\Gamma(\sum_k\gamma_k)}{\prod_k\Gamma(\gamma_k)}\right)^{T}\prod_{j}\frac{\prod_k\Gamma(n_{j,k}+\gamma_k)}{\Gamma(n_j+\sum_k\gamma_k)}, \tag{4.11}$$

$$p(\mathbf{z}|\alpha) = \left(\frac{\Gamma(\sum_j\alpha_j)}{\prod_j\Gamma(\alpha_j)}\right)^{D}\prod_{d}\frac{\prod_j\Gamma(n_{d,j}+\alpha_j)}{\Gamma(n_d+\sum_j\alpha_j)}, \tag{4.12}$$

Distribution a posteriori. La distribution *a posteriori* est obtenue en calculant la probabilité d'observer une thématique et une opinion (variables z et s) connaissant toutes les autres variables. Nous utilisons l'exposant "$-p$" pour représenter le sous-ensemble des données d'apprentissage qui exclut le terme à la position p du document courant. L'opinion et la thématique associées à ce terme sont notées s_p et z_p respectivement.

La distribution marginale de s_p et z_p est donnée par la formule suivante :

$$\begin{aligned} & p(s_p = k, z_p = j | \mathbf{w}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma) \\ = & \frac{p(s_p = k, z_p = j, w_p | \mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma)}{p(w_p | \mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma)} \\ = & \frac{p(\mathbf{w}, \mathbf{s}, \mathbf{z}, \alpha, \beta, \gamma)}{p(\mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma)} \cdot \frac{p(\mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma)}{p(w_p, \mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma)} \\ = & \frac{p(\mathbf{w}, \mathbf{s}, \mathbf{z} | \alpha, \beta, \gamma)}{p(\mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p} | \alpha, \beta, \gamma)} \cdot \frac{1}{w_p} \\ \propto & \frac{p(\mathbf{w}, \mathbf{s}, \mathbf{z} | \alpha, \beta, \gamma)}{p(\mathbf{w}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p} | \alpha, \beta, \gamma)} \end{aligned} \tag{4.13}$$

Le numérateur et le dénominateur de la fraction ci-dessus sont remplacés par le produit des trois facteurs déjà calculés dans le paragraphe précédent. Cela donne :

$$\begin{aligned} & p(s_p = k, z_p = j | \mathbf{w}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma) \\ \propto & \underbrace{\frac{p(\mathbf{w} | \mathbf{s}, \mathbf{z}, \beta)}{p(\mathbf{w}^{-p} | \mathbf{s}^{-p}, \mathbf{z}^{-p}, \beta)}}_{A} \cdot \frac{p(\mathbf{s} | \mathbf{z}, \gamma)}{p(\mathbf{s}^{-p} | \mathbf{z}^{-p}, \gamma)} \cdot \frac{p(\mathbf{z} | \alpha)}{p(\mathbf{z}^{-p} | \alpha)} \end{aligned} \tag{4.14}$$

Chacun de ces quatre facteurs peut être simplifié en utilisant la propriété de récursivité de la fonction Γ. Voici le déroulement de cette simplification sur le premier facteur, noté A :

$$\begin{aligned} A = & p(\mathbf{w} | \mathbf{s}, \mathbf{z}, \beta) \cdot p(\mathbf{w}^{-p} | \mathbf{s}^{-p}, \mathbf{z}^{-p}, \beta)^{-1} \\ = & \prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\prod_i \Gamma(n_{i,j,k}^{-p} + \beta + 1)}{\Gamma(n_{j,k}^{-p} + V\beta + 1)} \cdot \left[\prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\prod_i \Gamma(n_{i,j,k}^{-p} + \beta)}{\Gamma(n_{j,k}^{-p} + V\beta)} \right]^{-1} \\ = & \prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\prod_i (n_{w_p,j,k}^{-p} + \beta) \Gamma(n_{i,j,k}^{-p} + \beta)}{(n_{j,k}^{-p} + V\beta) \Gamma(n_{j,k}^{-p} + V\beta)} \cdot \prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\Gamma(n_{j,k}^{-p} + V\beta)}{\prod_i \Gamma(n_{i,j,k}^{-p} + \beta)} \\ = & \frac{n_{w_p,j,k}^{-p} + \beta}{n_{j,k}^{-p} + V\beta} \end{aligned} \tag{4.15}$$

Les deux autres facteurs de la probabilité jointe peuvent être obtenus de la même manière. Au final, la formule de calcul de la probabilité jointe devient :

$$\begin{aligned} & p(s_p = k, z_p = j | \mathbf{w}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \gamma) \\ \propto & \frac{n_{w_p,j,k}^{-p} + \beta}{n_{j,k}^{-p} + V\beta} \cdot \frac{n_{j,k}^{-p} + \gamma_k}{n_j^{-p} + \sum_k \gamma_k} \cdot \frac{n_{d,j}^{-p} + \alpha_j}{n_d^{-p} + \sum_j \alpha_j}. \end{aligned} \tag{4.16}$$

Les échantillons obtenus de cette chaîne de Markov sont ensuite utilisés pour estimer les distributions multinomiales φ, θ et π de la manière suivante

$$\varphi_{j,k,i} = \frac{n_{i,j,k} + \beta}{n_{j,k} + V\beta}, \theta_{d,j} = \frac{n_{d,j} + \alpha_j}{n_d + \sum_j \alpha_j}, \pi_{j,k} = \frac{n_{j,k} + \gamma_k}{n_j + \sum_k \gamma_k} \tag{4.17}$$

4.4.3 Intégration de la connaissance a priori

Dans le domaine des réseaux bayésiens, l'utilisation de la connaissance *a priori* est l'une des techniques couramment utilisée pour définir des distributions de probabilités [79]. Plusieurs travaux ont déjà montré l'efficacité de cette technique pour améliorer la détection des opinions avec des modèles probabilistes [37, 49, 52, 59, 60, 63, 64, 70, 115]. L'incorporation de cette connaissance a permis de "guider" le modèle afin d'extraire différentes polarités de la même thématique. A titre d'exemple, la non-utilisation de la connaissance *a priori* a montré que le modèle JST est très proche de l'aléatoire (60% de précision pour un problème de classement binaire de documents) alors que ce même modèle a atteint 74% avec l'incorporation d'un lexique composé seulement de 42 termes polarisés [63].

Nous adoptons ici une approche similaire à celle utilisée dans [63] afin d'intégrer de la connaissance sur les termes polarisés. L'idée consiste à imposer des probabilités d'affectation d'un terme aux polarités de l'opinion. Cette probabilité, pour la plupart des termes, est uniforme (aucune connaissance *a priori* sur la polarité du terme). Seulement une petite fraction des termes sont clairement polarisés, par exemple "bon", "mauvais", etc. La définition d'une distribution *a priori* pour ces termes permet de forcer le modèle à tirer une polarité d'opinion similaire à leur polarité *a priori*, par exemple le terme "bon" est toujours associé à la polarité positive.

La connaissance *a priori* se présente sous forme d'un lexique (liste de termes) dont on connaît la polarité. Cette connaissance est incorporée dans la phase d'apprentissage d'inférence au moment de tirer une opinion pour un terme (voir algorithme 1 ci-dessous). Ainsi, si le terme se trouve dans le lexique, cette opinion est tirée suivant la distribution qui lui est associée dans le lexique. Sinon, l'opinion est tirée sur la base d'une distribution uniforme, en utilisant l'équation 4.16.

Dans la pratique, l'obtention de ce type de lexiques n'est pas difficile ; plusieurs lexiques pour différentes langues ont déjà été développés pour des finalité d'analyse d'opinions (cf. section 3.2.3).

4.4.4 Algorithme d'inférence.

La procédure complète d'apprentissage par échantillonnage de Gibbs est donnée dans l'Algorithme 1.

Algorithm 1 Inférence pour le modèle TS

Require: α, β, γ, T, un lexique $\mathbb{L}$

1: Initialiser aléatoirement les matrices Φ, Θ, Π.
2: **for** iteration $c = 1$ **to** *nbGibbsIterations* **do**
3: **for** document $d = 1$ **to** D **do**
4: **for** $p = 1$ **to** n_d **do**
5: Exclure le terme w_p du document d et mettre à jour les variables de comptage
6: Tirer une thématique et une opinions pour le terme w_p en utilisant l'équation 4.16
7: **if** le terme w_p est dans le lexique $\mathbb{L}$ **then**
8: Changer l'opinion de w_p pour qu'elle corresponde à celle du lexique $\mathbb{L}$
9: **end if**
10: Mettre à jour les variables de comptage avec les nouvelles données
11: **end for**
12: **end for**
13: **end for**
14: Mettre à jour les matrices Φ, Θ, Π en utilisant les équations 4.17

Complexité algorithmique. Soit W le nombre de termes de tous les documents du corpus ($W = \sum_{d \in D} n_d$). La complexité algorithmique de tirer une thématique et une opinion pour un terme (ligne 6 de l'Algorithme 1) est $O(S \cdot T)$. Par conséquent, la complexité algorithmique de chaque itération de la procédure d'échantillonnage de Gibbs est $O(W \cdot S \cdot T)$. A noter que, pour des raisons de simplicité, la complexité algorithmique pour le tirage d'une multinomiale (dérivée de la génération automatique de nombres aléatoires) est supposée être $O(1)$.

4.5 Expérimentations

Dans cette section, nous présentons les jeux de données, la méthodologie d'évaluation ainsi que les résultats obtenus avec le modèle TS. Nous avons choisi de comparer le modèle TS à deux autres modèles de la littérature : JST et ASUM. Le modèle JST est un des travaux de référence dans ce domaine. Le modèle ASUM a montré de bonnes performances grâce à la prise en compte de la séquentialité des termes (les termes de la même phrase appartiennent à la même thématique). De plus, ces deux modèles ont des codes sources disponibles, ce qui facilite la reproductibilité des résultats.

4.5.1 Données et paramètres

Jeux de données. Pour tester le modèle TS, nous utilisons deux corpus de critiques collectés sur la plateforme de vente en ligne Amazon [1] : un corpus anglophone MDSen et un corpus francophone MDSfr. Le corpus MDSen a déjà été utilisé dans [10] pour l'analyse d'opinions multi-domaine. Chaque document du corpus est une critique d'un produit. Ainsi, pour l'évaluation, nous avons ré-annoté ces critiques selon deux dimensions : l'opinion et la thématique. L'opinion est automatiquement déduite à partir de l'évaluation des utilisateurs (1 étoile pour le négatif, 5 étoiles pour le positif). Le reste des documents est écarté du corpus. Pour la thématique, nous avons annoté les critiques de produits de la même catégorie (dans le sens d'Amazon) avec la même thématique. Par exemple, des critiques sur les produits "rasoir", "sèche-cheveux", "parfum" sont annotés par la même thématique qui correspond à la catégorie "beauté et parfums".

Le corpus MDSfr a été collecté manuellement sur la plateforme Amazon France [2]. Nous avons annoté les critiques sur les deux dimensions (thématique et opinion) de la même manière que MDSen. Nous avons effectué comme prétraitements la racinisation (*stemming*), la suppression de mots outils (*stopwords*) et des valeurs numériques. Le tableau 4.2 décrit ces deux corpus après ces prétraitements. L'avant-dernière colonne donne le nombre de thématiques majoritairement positives/négatives. Une thématique est majoritairement positive (respectivement négative) s'il y a plus de documents positifs (respectivement négatifs) affectés à cette thématique.

Paramètres. Les paramètres des trois modèles utilisés, TS, JST et ASUM sont donnés dans le tableau 4.3. Pour fixer le paramètre α, nous avons adopté le même choix de la littérature ($\alpha = \frac{50}{T}$) [63]. Le paramètre β a été empiriquement fixé à $\frac{1}{T}$. En effet, des expérimentations ont montré la pertinence de ce choix. Concernant le paramètre γ, il a été empiriquement fixé de manière à optimiser le score de chaque modèle. Dans la section 4.5.4, nous présentons une méthode pour fixer automatiquement ce paramètre.

4.5.2 Méthodologie d'évaluation

Comme cela a été souligné dans la section 4.3, les modèles probabilistes pour l'extraction conjointe des thématiques et des opinions ont été évalués à partir de leur capacité à prédire la bonne classe d'opinion au niveau du document. Or, ceci n'est pas l'objectif premier de ce type de modèles. Pour nous, de tels modèles doivent être évalués au moins sur la dimension de la

1. http ://www.amazon.com
2. http ://www.amazon.fr

Corpus	Langue	D	V	#thém. pos./nég.	thématiques
MDSfr	Français	10 668	12 773	9/8	animalerie, auto moto, bagages, beauté et parfums, bijoux, bricolage, bureau, chaussures, cuisine, électroménager, informatique, instruments de musique, photo, puériculture, sports et loisirs, téléphones, vêtements et accessoires
MDSen	Anglais	27 065	42 010	12/12	apparel, automotive, beauty, books, camera and photo, cell phones and service, computer and video games, dvd, electronics, gourmet food, grocery, health and personal care, jewelry and watches, kitchen and housewares, magazines, music, musical instruments, office products, outdoor living, software, sports and outdoors, tools and hardware, toys and games, video

TABLE 4.2 – Statistiques des corpus MDSfr et MDSen.

thématique, i.e., capacité du modèle à prédire la bonne opinion au niveau de la thématique. Pour cela, nous nous appuyons sur un corpus annoté afin de construire une vérité terrain par rapport à laquelle sera évalué notre modèle ainsi que deux autres modèles de la littérature : JST et ASUM. Les raisons du choix de ces deux modèles ont déjà été données dans la section 4.5.

Construction de la vérité terrain. Les documents des corpus MDSfr et MDSen sont annotés avec la thématique (catégorie du produit) et l'opinion (positive ou négative). Nous nous appuyons sur cette annotation afin de construire une vérité terrain concernant les opinions relatives aux thématiques de la manière suivante :

1. Le corpus de données est d'abord divisé en T sous-ensembles où chaque sous-ensemble $\mathbb{D}_j$ contient les documents annotés avec la thématique d'indice j.
2. Chaque thématique est ensuite annotée par l'une des deux polarités, positive ou négative, en se basant sur la proportion des documents positifs/négatifs qu'elle contient (vote majoritaire).

Corpus	Modèle	Paramètres			
		T	α	β	(γ_+, γ_-)
MDSfr	TS	17	2.94	0.06	(0.1, 20)
	JST	17	2.94	0.06	(0.1, 2)
	ASUM	17	2.94	0.06	(0.1, 10)
MDSen	TS	24	2.08	0.04	(0.1, 200)
	JST	24	2.08	0.04	(0.1, 2)
	ASUM	24	2.08	0.04	(0.1, 50)

TABLE 4.3 – Paramètres des modèles utilisés pour l'évaluation.

Calcul du taux de succès. Cette vérité terrain est ensuite utilisée pour calculer une erreur de classement classique (nombre de thématiques mal-classées par le modèle). Nous utilisons ici le taux de succès qui est égal à la proportion des thématiques bien-classées (nombre de thématiques bien classées divisé par le nombre de thémmatiques total T). La procédure d'évaluation se déroule donc en deux temps :

1. Chacune des thématiques réelles est associée à une thématique estimée. Cette étape peut être réalisée manuellement.
2. Les classes d'opinions réelles sont obtenues à partir des annotations décrites plus haut (construction de la vérité terrain). Les classes d'opinions estimées sont obtenues pour chaque thématique z par la maximisation des distributions π (une thématique est annotée avec la classe d'opinion m si $m = \arg\max_s \pi_{z,s}$).

4.5.3 Résultats

Résultats quantitatifs. Selon notre approche d'évaluation, un bon modèle est un modèle qui prédit la bonne classe d'opinion au niveau de la thématique. Nous utilisons ce critère pour comparer le modèle TS aux modèles JST et ASUM. Afin d'éliminer le biais du à l'initialisation aléatoire de ces modèles, nous effectuons cinq exécutions pour chaque modèle et nous reportons la moyenne et l'écart-type du taux de succès sur le tableau 4.4.

Comme le montre ces résultats, la performance du modèle TS est supérieure à celle des modèles JST et ASUM en terme de prédiction de l'opinion au niveau de la thématique. Sur les deux corpus MDSen et MDSfr, le meilleur taux de succès a été obtenu par le modèle TS suivie de ASUM. En plus de cela, parmi les trois modèles testés, TS était le plus robuste à l'initialisation aléatoire par rapport aux deux autres modèles comme en témoigne les faibles valeurs de l'écart-type.

Corpus	Modèle	Taux de succès moyenne	Ecart-type
	TS	**0.765**	**0.042**
MDSfr	JST	0.541	0.049
	ASUM	0.718	0.049
	TS	**0.750**	**0.029**
MDSen	JST	0.600	0.037
	ASUM	0.667	0.029

TABLE 4.4 – Résultats de prédiction de l'opinion au niveau de la thématique. Moyenne et écart-type pour 5 initialisations aléatoires.

Résultats qualitatifs. Dans le tableau 4.5, nous présentons une sélection de thématiques extraites avec le modèle TS sur les deux corpus MDSfr et MDSen. Les lignes π réel et π est. représentent la distribution réelle, respectivement estimée, des thématiques sur les polarités d'opinion. Remarquons qu'aucun travail manuel n'est nécessaire pour visualiser ce résultat mis à part le nommage (en gras) qui consiste à assigner les noms aux thématiques extraites. Le modèle TS permet de produire pour chaque thématique deux descriptions correspondant aux deux polarités de l'opinion.

Sur la base de ces résultats, nous pouvons faire deux remarques concernant la qualité des thématiques et l'association des thématiques aux opinions. Les thématiques extraites sont clairement homogènes et porteuses de sens. En effet, parmi les termes probables de chaque thématique, la plupart sont sémantiquement liés et réfèrent à la même thématique. Par exemple, sur le corpus francophone MDSfr, la thématique relative à la catégorie "chaussures" est décrite par les termes "taille", "tissu", "lavage", "paire", "pied", etc. De même, la thématique relative à la catégorie de produits de beauté est décrite par les termes "savon", "odeur", "peau", "douche", etc.

De plus, nous remarquons l'apparition de termes polarisés (présents dans le lexique) parmi les termes probables. Dans la majorité des thématiques représentées ici, ces termes sont correctement répartis entre les deux polarités de l'opinion. En effet, les termes positifs (en vert souligné) sont plus probables dans les descriptions positive des thématiques et *vice versa*, les termes négatifs (en rouge italique) sont plus probables dans les descriptions négatives.

L'apparition des termes polarisés conduit à un autre résultat utile concernant la variation de la terminologie suivant le domaine. En effet, il y a des termes qui sont spécifiques à des domaines particuliers. Par exemple, le terme "sèche" est beaucoup plus fréquent dans la thématique "beauté et parfums" car c'est un terme du domaine. Le terme "encombrant" est beaucoup plus utilisé dans la thématique "puériculture".

Les mêmes constatations peuvent être faites sur le corpus anglophone MDSen (tableau 4.5, haut). Les thématiques représentées sont significatives et sémantiquement homogènes. Les thématiques sont décrites sous les deux po-

Thématique	**chaussures**		**beauté et parfum**		**puériculture**	
	positif	*négatif*	*positif*	*négatif*	*positif*	*négatif*
Termes probables	couleur	taille	huile	savon	pratique	*encombrant*
	pinceau	*trop*	peau	*sèche*	facile	*trop*
	chaussures	coup	utiliser	odeur	ranger	*mal*
	joli	noir	produit	*détruire*	enfant	fermeture
	marque	*déçu*	douche	*poubelle*	poussette	*compliqué*
	qualité	tissu	hydrater	noir	intérieur	pliage
	paire	lavage	efficace	texture	biberon	démonter
	pied	*dommage*	gel	marseille	transport	bébé
π réel	0.61	0.39	0.62	0.38	0.43	0.57
π est.	0.82	0.18	0.46	0.54	0.11	0.89

Thématique	**video**		**toys and games**		**music**	
	positif	*négatif*	*positif*	*négatif*	*positif*	*négatif*
Termes probables	movie	story	kid	old	song	guitar
	film	*bad*	love	toy	album	sing
	good	performance	play	year	music	*boring*
	scene	*boring*	fun	daughter	good	blue
	character	*horror*	christmas	*disappointed*	rock	*miss*
	actor	made	gift	baby	love	*bad*
	great	*worst*	enjoy	age	fan	remix
	play	*stupid*	learn	*frustrated*	original	*lack*
π réel	0.40	0.60	0.20	0.80	0.23	0.77
π est.	0.38	0.62	0.53	0.47	0.44	0.56

TABLE 4.5 – Termes probables pour une sélection de thématiques extraites avec le modèle TS sur le corpus MDSen (haut) et MDSfr (bas). Les termes du lexique sont coloriés en vert/souligné (positifs) et rouge/italique (négatifs).

larités d'opinion avec des termes polarisés et correctement répartis.

Cette évaluation qualitative a permis de confirmer que l'intégration de la dimension de l'opinion n'a pas remis en cause l'objectif premier du modèle TS, à savoir l'extraction de thématiques homogènes. En effet, le modèle TS permet non seulement d'extraire des thématiques homogènes et sémantiquement significatives mais aussi de les caractériser avec des distributions sur les opinions.

4.5.4 Fixer automatiquement le paramètre γ du modèle TS

Il a été montré par Hatzivassiloglou and McKeown que les termes de polarité positive sont plus fréquemment utilisés que les termes à polarité négative dans la langue anglaise [43]. Cette constatation est également valable pour le corpus francophone MDSfr que nous utilisons pour évaluer le modèle TS. En

examinant l'algorithme d'échantillonnage de Gibbs (Algorithme 1), nous remarquons que ce phénomène a un impact sur le déroulement de l'algorithme. En particulier, c'est sur la base du nombre de termes dans chaque polarité que l'affectation des termes aux polarités de l'opinion est décidée (équation 4.16). La distribution non uniforme des termes d'opinion (plus de termes positifs que de négatifs) induit une affectation non uniforme des termes du corpus aux polarités d'opinions. Les termes ont tendance à être affectés plus souvent à une polarité positive. Par conséquent, la plus grande partie des thématiques est affectée à la polarité positive également.

La technique la plus courante pour corriger ce biais est de choisir des hyperparamètres asymétriques [49, 59, 63, 64]. Il s'agit ici du paramètre γ qui contrôle l'affectation des termes aux polarités de l'opinion. La technique consiste à mettre un poids plus important sur la polarité négative, par exemple $\gamma_- = 0.5$ et $\gamma_+ = 0.01$. Cette stratégie a permis de rééquilibrer la distribution des termes sur les polarités de l'opinion. Cependant, elle a deux limites majeures que nous résumons comme suit :

1. La meilleure valeur du paramètre asymétrique est difficile à trouver car elle dépend directement des données. Généralement, celle-ci est fixée de manière empirique, pour chaque corpus de données, sur la base de plusieurs expérimentations.
2. Une fois γ fixé pour un corpus donné, cette valeur est seulement valable pour un nombre d'itérations donné (de l'algorithme d'échantillonnage de Gibbs). Par exemple, si γ a été fixé sur la base de 100 itérations, la valeur trouvée n'est pas nécessairement valable pour 200 itérations car l'effet de l'*a priori* négatif sera trop important et finit par déséquilibrer le modèle.

Notre proposition. Pour résoudre ce problème, nous proposons de mettre à jour la valeur de γ de manière dynamique, i.e. à chaque itération d'échantillonnage. Soient W_+ respectivement W_- le nombre de termes affectés à la polarité positive, respectivement négative, par l'algorithme d'échantillonnage de Gibbs au terme de l'itération c. Trois cas de figure sont possibles : soit $W_+ > W_-$, soit $W_+ < W_-$, soit $W_+ = W_-$. L'idée consiste à rendre, si ce n'est déjà le cas, la distribution des termes équilibrée sur les deux polarités de l'opinion ($W_+ = W_-$). Pour rappel, W_+ et W_- sont additionnées avec les valeurs des hyperparamètres γ_+ et γ_- (cf. équation 4.16).

Afin d'atteindre une situation d'équilibre, l'*a priori* sur la classe majoritaire doit être diminué et celui sur la classe minoritaire augmenté. Concrètement, dans l'itération suivante $c+1$, les valeurs des paramètres γ_+ et γ_- sont mises à jour de manière à atteindre l'équilibre. Les nouvelles valeurs de ces paramètres correspondent aux nombres de termes positifs, respectivement négatifs, qu'on doit rajouter au corpus afin d'atteindre $W_+ = W_-$. Ainsi :

$$\gamma_+ = \begin{cases} \frac{W_- - W_+}{T}, & \text{si } W_- > W_+ \\ \epsilon, & \text{sinon} \end{cases} \tag{4.18}$$

et

$$\gamma_- = \begin{cases} \frac{W_+ - W_-}{T}, & \text{si } W_+ > W_- \\ \epsilon, & \text{sinon} \end{cases} \tag{4.19}$$

Par exemple, si $W_+ > W_-$, γ_- sera augmenté à l'itération suivante avec le nombre moyen de termes qu'il faut rajouter artificiellement à chaque thématique afin d'atteindre la situation d'équilibre.

Le paramètre ϵ peut être remplacé par une très petite valeur par rapport à la valeur de la polarité opposée. Dans nos expérimentations, celui-ci a été fixé à 0.01.

Expérimentations. Notre méthode d'estimation du paramètre γ est évaluée en la comparant par rapport à la méthode à base de la maximisation de vraisemblance ML (*Maximum Likelihood*). Cette dernière a déjà été utilisée dans de nombreux modèles probabilistes pour fixer les paramètres d'*a priori* caractérisant les distributions des thématiques sur le vocabulaire α comme dans [64, 113] ou les distributions des documents sur les thématiques β comme dans [113]. Dans [113], la méthode à base de ML s'est montrée efficace pour fixer les paramètres du modèle LDA, notamment le paramètre α. Ici, nous comparons notre méthode avec la méthode ML. Cette comparaison est basée sur le critère décrit dans la section 4.5.2 (prédiction de l'opinion au niveau de la thématique).

Le principe de la méthode à base de ML est de partir d'une situation du modèle où chaque terme du corpus est affecté à une thématique, éventuellement à une opinion, puis de trouver les meilleurs valeurs des paramètres de Dirichlet qui auraient engendré cette affectation. Pour cela, plusieurs méthodes mathématiques peuvent être utilisées, comme la descente du gradient, la méthode de Newton-Raphson ou l'itération à point fixe. Une introduction à l'estimation des paramètres de Dirichlet peut être consultée dans [46]. Pour nos expérimentations, nous avons implémenté la méthode d'itération à point fixe décrite dans [76].

Comme le montre le tableau 4.6, notre méthode de paramétrage du modèle TS est plus performante que la méthode ML. Elle permet une meilleure prédiction de l'opinion au niveau de la thématique.

4.6 Discussion

Dans cette section, nous avons proposé le modèle TS pour résoudre le problème de modélisation conjointe des thématiques et des opinions. Notre

Corpus	Méthode	Taux de succès moyenne	Ecart-type
MDSfr	**Notre méthode**	**0.765**	**0.042**
	Méthode à base de ML	0.588	0.049
MDSen	**Notre méthode**	**0.750**	**0.029**
	Méthode à base de ML	0.542	0.029

TABLE 4.6 – Résultats obtenus avec deux méthodes pour fixer le paramètre γ du modèle TS : notre méthode et la méthode basée sur la maximisation de vraisemblance (ML). Moyenne et écart-type basés sur 5 initialisations aléatoires.

approche se distingue des travaux précédents dans ce domaine principalement par deux points :

- Chaque thématique est caractérisée par plusieurs distributions sur le vocabulaire, une pour chaque polarité d'opinion. Les expérimentations conduites sur le modèle TS ont montré que ces différentes thématiques, tout en étant associées aux polarités de l'opinion, préservent l'homogénéité et la sémantique globale de la thématique. L'autre avantage de cette approche est de pouvoir se passer du post-traitement afin de faire correspondre les différentes perspectives de la même thématique. Cela permet au modèle TS de mieux appréhender les associations thématiques-opinions et a conduit à de meilleures performances en terme de prédiction de l'opinion au niveau de la thématique par rapport aux modèles basés sur le post-traitement.
- La portée des opinions relatives aux thématiques est un aspect qui n'a pas été suffisamment traité dans les modèles existants. Dans TS, la distribution d'une thématique sur les polarités d'opinion est une connaissance qui est extraite à partir de tous les documents du corpus en une seule fois. Cette fonctionnalité permet d'obtenir une vue d'ensemble des associations thématiques-opinions qui peut se révéler plus utile qu'au niveau du document.

La méthode de mise à jour automatique des paramètres du modèle TS que nous avons proposée est d'une grande utilité pour une utilisation facile du modèle et à plus forte raison pour l'industrialisation efficace de ces travaux. Cependant, cette méthode repose sur l'hypothèse suivante : les deux polarités d'opinions opposées (positive et négative) sont uniformément représentées par les termes du corpus. Cela permet de maintenir l'équilibre du modèle (cf. section 4.5.4) mais le rend plus propice aux biais quand l'hypothèse n'est pas vérifiée, par exemple dans un contexte où le corpus est composé uniquement de documents positifs, l'utilisation de cette méthode peut biaiser le déroulement de l'apprentissage en "forçant" l'apparition des termes négatifs. Cependant, nous pouvons supposer que ce problème ne se posera pas dans un contexte de

données réelles car des expérimentations ont montré que même dans des situations extrêmes le modèle reste efficace pour extraire les proportions d'opinions.

Notre méthodologie d'évaluation est basée sur le critère de prédiction des opinions au niveau de la thématique. Notre approche consiste à considérer qu'un bon modèle devrait arriver à bien estimer l'opinion au niveau de la thématique. La vérité terrain que nous avons construite pour réaliser cette évaluation est basée sur le vote majoritaire, i.e., une thématique avec des documents majoritairement positifs est annotée définitivement positive. Or, cette approche peut amener à omettre le caractère flou de cette annotation. Dans le chapitre suivant, nous proposons une autre approche d'évaluation qui tient compte de ce phénomène et nous l'utilisons pour évaluer un autre modèle orienté vers l'analyse dynamique des thématiques et des opinions.

Chapitre 5

Thématiques et Opinions : Modélisation Conjointe et Dynamique

Résumé. *Après avoir traité la modélisation conjointe des thématiques et des opinions dans le chapitre précédent, nous nous intéressons désormais à la dynamique de celles-ci. Il s'agit d'extraire l'évolution de la quantité de données liées à chaque thématique/opinion. Pour cela, nous proposons une extension du modèle TS afin d'inclure la dimension temporelle. Nous démontrons la performance de notre modèle par rapport aux modèles de l'état de l'art à travers une évaluation adaptée en utilisant des jeux de données composés de critiques en ligne et d'articles de presse.*

Sommaire

5.1 Introduction

Les données issues du Web sont de nature évolutive car le contenu et la structure de ces données changent dans le temps. Prenons l'exemple de Twitter. Une étude récente a montré l'augmentation constante de certains types de données comme les spams et le contenu indésirable [66]. Cet exemple illustre bien le phénomène de l'évolution des données en terme de quantité. D'autres types de changement peuvent survenir dans le temps comme l'apparition ou la disparition de liens entre les utilisateurs, le passage aux applications mobiles, etc. Tous ces changements influencent les contenus créés par les utilisateurs et, plus généralement, leur comportement [102]. Ainsi, un utilisateur peut devenir plus ou moins actif, plus ou moins productif, etc.

L'évolution du contenu Web a été étudiée selon plusieurs axes dont celui de l'évolution des thématiques. Le problème consiste à modéliser le changement des thématiques à travers le temps. Plusieurs types de changement peuvent être considérés, tels que le changement dans la quantité de données liée à la thématique, le changement des termes décrivant la thématique, le changement dans les associations entre les thématiques, etc.

Ici, nous nous intéressons à une problématique nouvelle : l'évolution des thématiques conjointement aux opinions. En particulier, il s'agit de ce que nous appelons l'évolution quantitative, c'est-à-dire le changement en terme de quantité de données liées à une thématique et une opinion. En effet, le problème n'a été abordé que très peu dans la littérature. La plupart des modèles de la littérature se base sur un post-traitement comme moyen de modéliser l'évolution temporelle. Ici, nous proposons un modèle conjoint pour l'extraction des thématiques et des opinions ainsi que leur évolution dans le temps.

Notre proposition consiste à étendre le modèle TS précédemment décrit afin d'inclure l'information temporelle. Nous proposons une évaluation adaptée basée sur le calcul d'une distance entre l'estimation obtenue par le modèle et une vérité terrain. Nous montrons la performance de notre modèle par rapport aux modèles de l'état de l'art qui n'ont pas vocation initiale l'analyse de l'évolution temporelle mais qui peuvent être adaptés afin d'y parvenir.

La section suivante présente un état de l'art. La section 5.3 présente notre contribution. La section 5.4 présente les expérimentations réalisées pour tester notre modèle. Enfin, la section 5.5 est consacrée à une discussion des résultats obtenus et quelques perspectives.

5.2 Etat de l'art

De manière générale, les thématiques évoluent selon deux dimensions : quantitativement et qualitativement. L'évolution quantitative concerne le vo-

lume de données relatives à une certaine thématique, alors que l'évolution qualitative concerne toutes les autres caractéristiques de la thématique (e.g., vocabulaire, corrélation avec les autres thématiques). Dans la suite de cet état de l'art, nous traitons ces deux types d'évolution séparément.

5.2.1 Evolution qualitative

L'évolution qualitative ne s'intéresse pas à la quantité de données traitant une thématique mais à la thématique elle-même selon plusieurs axes (vocabulaire, corrélation aux autres thématiques, etc.). Considérer ici toutes ces caractéristiques risque d'élargir le spectre de cet état de l'art et de l'éloigner du point principal, à savoir l'évolution quantitative des thématiques. Pour cette raison, notre étude sera centrée sur l'aspect le plus traité dans la littérature, à savoir celui de l'évolution du vocabulaire.

Définition 5 (Evolution qualitative des thématiques) *Le problème de modélisation de l'évolution qualitative des thématique consiste à trouver une fonction* $f_{qual.evol} : \mathbb{Z} \times \mathbb{Q} \rightarrow \mathbb{F}$*, où* $\mathbb{Q}$ *représente l'axe temporel et* $\mathbb{F}$ *l'ensemble de toutes les combinaisons linéaires sur le vocabulaire. Cette fonction associe à chaque paire thématique-étiquette temporelle, une combinaison linéaire de termes* $\varphi_{z,t} \in \mathbb{F}$*.*

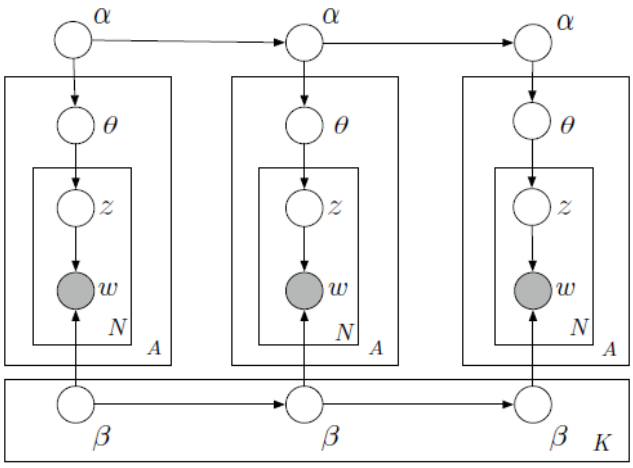


FIGURE 5.2.1 – Modèle graphique de DTM.

DTM (*Dynamic Topic Model*) [7] est un modèle à base de LDA qui permet de capturer ce type d'évolution temporelle. Les documents de la collection sont d'abord groupés selon l'étiquette temporelle. Ensuite, un modèle de type LDA est appris sur chaque groupe de documents. Un modèle appris à l'étiquette de temps t est obtenu en faisant évoluer le modèle à l'instant $t-1$. La relation

entre ces deux modèles se concrétise par la liaison des paramètres α, β entre les deux modèles (cf. figure 5.2.1). Ainsi, à l'instant t, les paramètres α_t et β_t sont tirées à partir de distributions Gaussiennes.

$$\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \sigma^2 I), \quad \beta_t \sim \mathcal{N}(\beta_{t-1}, \delta^2 I). \tag{5.1}$$

TopicMonitor [36] est un autre modèle pour l'évolution du vocabulaire. L'évolution est capturée à l'aide d'une "fenêtre" qui glisse dans le flux de documents. Ainsi, les termes apportés par les nouveaux documents sont rajoutés dans le vocabulaire. Ce modèle permet de modéliser non seulement le changement dans les distributions sur le vocabulaire mais aussi l'évolution du vocabulaire lui-même.

5.2.2 Evolution quantitative

Soit $\mathbb{Z}$ un ensemble de thématiques donné, et $t \in \mathbb{T}$ une variable temporelle qui peut être discrète (par exemple jours, mois, années) ou continue (par exemple $\mathbb{T} = \mathbb{R}^+$). Nous entendons par évolution quantitative d'une thématique la variation en terme de quantité de données associées à cette thématique suivant l'axe temporel.

Définition 6 (Evolution quantitative des thématiques) *Le problème de modélisation de l'évolution quantitative des thématique équivaut à trouver une fonction $f_{quant.evol} : \mathbb{Z} \times \mathbb{T} \rightarrow \mathbb{R}^+$ qui associe à chaque paire thématique-étiquette temporelle, un score positif qui reflète la quantité de données traitant la thématique z à l'instant t.*

Un travail majeur dans cette catégorie est celui de Wang et McCallum [116] avec le modèle TOT (*Topics over Time*). TOT est un modèle probabiliste à base de LDA pour l'évolution quantitative des thématiques dans le temps. La méthode LDA a été étendue avec une nouvelle variable observée t pour capturer l'évolution temporelle. Ainsi, le processus génératif (cf. figure 5.2.2) permet de générer simultanément les mots et les étiquettes temporelles pour un document. Alors que les mots sont générés par une distribution multinomiale comme dans LDA, les étiquettes temporelles sont générées par une loi continue de type Béta.

Un des points forts du modèle TOT réside dans la prise en compte de l'information temporelle pour extraire les thématiques. En effet, les thématiques ne sont pas seulement définies comme des groupes de termes qui co-occurrent mais aussi qui co-occurrent au même instant (même étiquette temporelle). Cela aide à améliorer l'homogénéité des thématiques extraites et de distinguer deux thématiques qui partagent beaucoup de termes mais qui, dans le temps, sont éloignées l'une de l'autre. Un exemple est donné où TOT réussit à séparer deux

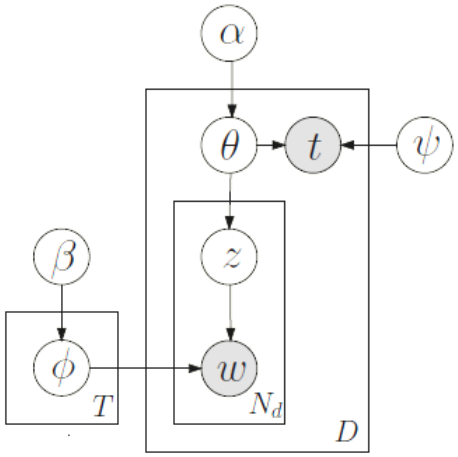


FIGURE 5.2.2 – Modèle graphique de TOT.

thématiques : "guerre américo-mexicaine" (1846-1848) et "première guerre mondiale" (1914-1918). Ces thématiques sont très similaires dans leur vocabulaire mais les documents qui en parlent sont situés dans deux périodes éloignées.

Dans [39], un post-traitement est utilisé pour estimer la popularité d'un ensemble de thématiques extraites avec le modèle LDA à partir d'un corpus d'articles scientifiques (e.g., "bases de données", "apprentissage automatique", "réseaux de neurones"). Un travail similaire a été proposé dans [73] pour analyser la popularité des genres cinématographiques à partir d'un corpus de critiques de films. Dans les deux travaux, le post-traitement consiste à calculer la proportion de mots ou de documents affectés à chacune des thématiques puis à projeter cette information sur l'axe temporel. Contrairement au modèle TOT, les corrélations temporelles entres les thématiques ne sont prises en compte.

Ces travaux ne modélisent pas les opinions liées aux thématiques. A notre connaissance, il y a seulement deux travaux qui se sont intéressés à cette problématique [70, 131]. Dans [70], le modèle TSM précédemment utilisé pour l'extraction conjointe des thématiques et des opinions a été utilisé pour modéliser l'évolution de celles-ci en le combinant à une méthode de post-traitement basée sur le comptage de termes. Dans [131], les auteurs, quand à eux, ont intégré une variable temporelle au modèle TSM leur permettant d'obtenir l'évolution temporelle des thématiques sans post-traitement des résultats. Cependant, ce modèle est toujours basé sur PLSA et présente plusieurs problèmes, notamment une tendance au sur-apprentissage et sa grande complexité algorithmique par rapport au modèle LDA (cf. section 2.3.1).

Afin de modéliser l'évolution des thématiques et des opinions qui y sont liées, nous proposons un nouveau modèle qui combine trois variables correspondant aux trois aspects du texte : thématique, opinion et temps. Notre

modèle fait usage des paramètres de Dirichlet (comme dans le modèle LDA). Ces paramètres de Dirichlet sont particulièrement utiles pour la modélisation conjointe thématiques-opinions car, comme nous l'avons déjà montré dans le chapitre précédent (section 4.5.4), un bon choix de ces paramètres a permis d'améliorer la performance du modèle TS.

5.3 Contribution : le modèle TTS (*Time-aware Topic-Sentiment model*)

Comme il a été déjà souligné dans la section 5.2, les données Web évoluent constamment. Nous nous intéressons ici à la modélisation de l'évolution quantitative (cf. Définition 6) des thématiques et des opinions qui s'y rapportent. Pour cela, nous adoptons une démarche similaire aux travaux de [116] où la variable temporelle observée (heure, date,..) est directement intégrée dans le modèle. En effet, l'information temporelle n'est généralement pas difficile à obtenir (date de parution d'un article de presse, date de création d'un tweet, date de publication d'un article scientifique, etc.). Nous nous appuyons sur cette information afin de fournir une analyse de l'évolution des thématiques et des opinions dans le temps.

Le modèle TTS (*Time-aware Topic-Sentiment model*) que nous proposons est une extension du modèle TS décrit dans la section précédente. Ce dernier a permis l'extraction conjointe des thématiques et des opinions relatives. Ici, nous proposons d'aller plus loin dans cette analyse et de modéliser l'évolution de ces relations à l'aide du modèle TTS comme l'illustre la figure 5.3.1.

5.3.1 Modèle graphique et processus génératif

Dans TTS, les thématiques et les opinions sont associées à une nouvelle variable temporelle discrète t (cf. figure 5.3.2). Par conséquent, la distribution ψ associée à la variable t est une distribution multinomiale.

Le processus génératif du modèle TTS permet de générer des termes pour des documents et pour chaque terme, il permet de générer une étiquette temporelle. Le processus génératif des termes et des étiquettes temporelles correspondant au modèle graphique de la figure 4.4.2 se déroule comme suit :

1. Tirer $T \times S$ multinomiales $\varphi_{z,s} \sim \text{Dirichlet}(\beta)$
2. Tirer $T \times S$ multinomiales $\psi_{z,s} \sim \text{Dirichlet}(\mu)$
3. Tirer T multinomiales $\pi_z \sim \text{Dirichlet}(\gamma)$
4. Pour chaque document d, tirer une multinomiale $\theta_d \sim \text{Dirichlet}(\alpha)$, puis pour chaque terme w_i dans d :
   (a) Tirer une thématique $z_i \sim \theta_d$

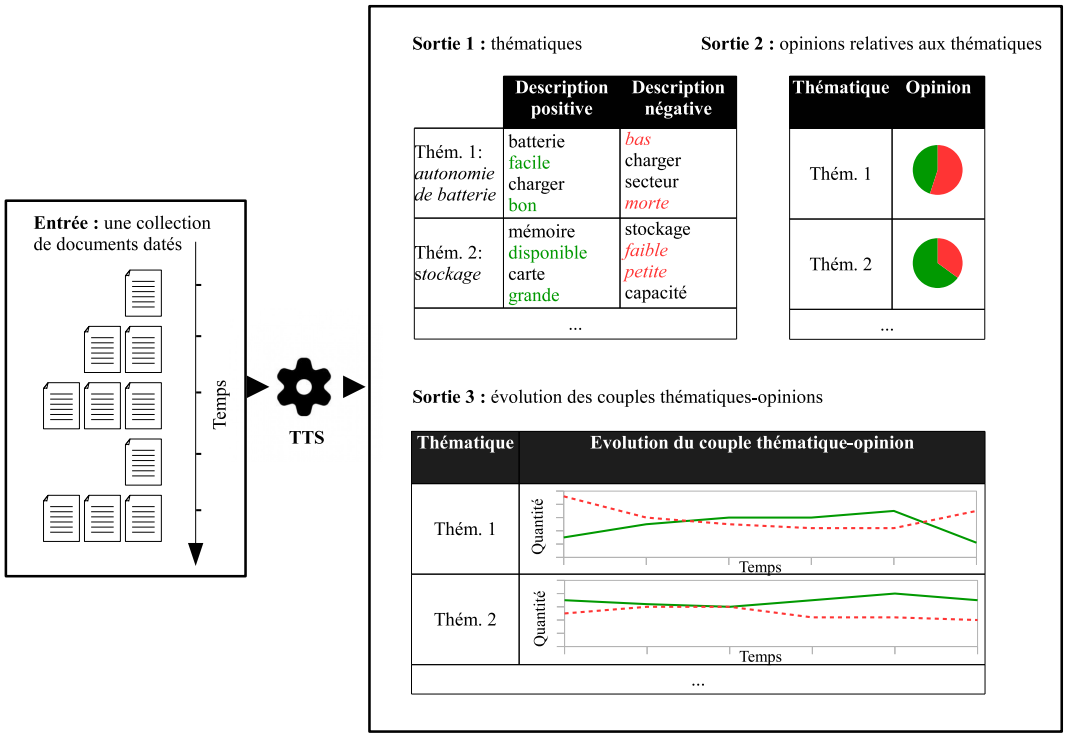


FIGURE 5.3.1 – Modélisation dynamique des thématiques et des opinions avec le modèle TTS.

(b) Tirer une opinion $s_i \sim \pi_{z_i}$

(c) Tirer un terme $w_i \sim \varphi_{z_i,s_i}$

(d) Tirer une étiquette temporelle $t_i \sim \psi_{z_i,s_i}$

Les opérations (c) et (d), étant indépendantes, elles peuvent être permutées sans aucun impact sur le processus.

5.3.2 Inférence

L'inférence consiste à estimer les variables latentes du modèle TTS (paramètres) φ, θ et ψ à partir des variables observées (w et t). Pour ce faire, nous utilisons la méthode d'échantillonnage de Gibbs (cf. section 4.4.2).

Comme pour le modèle TS, nous avons d'abord besoin de calculer la distribution jointe des termes, des thématiques et des opinions : $p(\mathbf{w}, \mathbf{s}, \mathbf{z}|\alpha, \beta, \gamma, \mu)$. Ensuite, nous déduisons la distribution *a posteriori* qui correspond à la probabilité de tirer une thématique et une opinion pour un terme donné. Nous nous appuyons sur les définitions de la loi de Dirichlet et des fonctions Γ et Δ données dans la section 4.4.2.

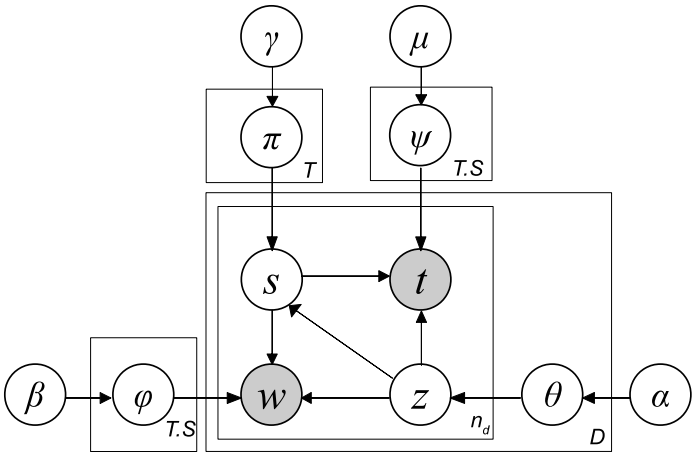


FIGURE 5.3.2 – Modèle graphique de TTS.

Distribution jointe. En utilisant la règle d'indépendance de Bayes, la probabilité jointe d'observer un terme, une thématique et une opinion peut être décomposée comme suit :

$$\begin{aligned} & p(\mathbf{w}, \mathbf{t}, \mathbf{s}, \mathbf{z} | \alpha, \beta, \gamma, \mu) \\ = \; & p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \beta) \cdot p(\mathbf{t}|\mathbf{s}, \mathbf{z}, \mu) \cdot p(\mathbf{s}|\mathbf{z}, \gamma) \cdot p(\mathbf{z}|\alpha). \end{aligned} \tag{5.2}$$

Les facteurs de la distributions jointe peuvent être calculés indépendamment les uns des autres. Pour obtenir le premier facteur, cherchons d'abord la distribution $p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \varphi)$. Celle-ci peut être directement calculée à partir de φ en multipliant les contributions provenant des termes et des opinions :

$$p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \varphi) = \prod_{j=1}^{T} \prod_{k=1}^{S} \prod_{i=1}^{V} \varphi_{j,k,i}^{n_{i,j,k}} \tag{5.3}$$

Les indices i, j, k et h sont utilisés pour indexer les termes, les thématiques, les opinions et les étiquettes temporelles respectivement.

Le premier facteur de l'équation 4.5 est obtenu en intégrant par rapport à φ :

$$p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \beta) = \int p(\mathbf{w}|\mathbf{s}, \mathbf{z}, \varphi) \cdot p(\varphi|\beta) \, \mathrm{d}\varphi \tag{5.4}$$

En utilisant la loi de probabilité de Dirichlet (équation 4.1) et l'équation 5.3 :

$$
\begin{aligned}
p(\mathbf{w}|\mathbf{s},\mathbf{z},\beta) &= \int \prod_{j=1}^{T}\prod_{k=1}^{S}\left(\prod_{i=1}^{V}\varphi_{j,k,i}^{n_{i,j,k}}\cdot\frac{1}{\Delta(\beta)}\prod_{i=1}^{V}\varphi_{j,k,i}^{\beta_i-1}\right)\mathrm{d}\varphi_{j,k} \\
&= \prod_{j=1}^{T}\prod_{k=1}^{S}\left(\frac{1}{\Delta(\beta)}\cdot\int\prod_{i=1}^{V}\varphi_{j,k,i}^{n_{i,j,k}+\beta_i-1}\,\mathrm{d}\varphi_{j,k}\right) \qquad (5.5)
\end{aligned}
$$

En réécrivant l'intégrale de Dirichlet en fonction de Δ (équation 4.4) :

$$
p(\mathbf{w}|\mathbf{s},\mathbf{z},\beta) = \prod_{j=1}^{T}\prod_{k=1}^{S}\frac{1}{\Delta(\beta)}\cdot\Delta(n_{i,j,k}+\beta) \qquad (5.6)
$$

En remplaçant maintenant la fonction Δ par sa première définition utilisant la fonction Γ (équation 4.2) :

$$
p(\mathbf{w}|\mathbf{s},\mathbf{z},\beta) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^{T\cdot S}\prod_{j}\prod_{k}\frac{\prod_i\Gamma(n_{i,j,k}+\beta)}{\Gamma(n_{j,k}+V\beta)}, \qquad (5.7)
$$

Le duxième facteur de l'équation 5.2 est obtenu de manière identique en intégrant par rapport à ψ :

$$
p(\mathbf{t}|\mathbf{s},\mathbf{z},\mu) = \left(\frac{\Gamma(H\mu)}{\Gamma(\mu)^H}\right)^{T\cdot S}\prod_{j}\prod_{k}\frac{\prod_h\Gamma(n_{j,k,h}+\mu)}{\Gamma(n_{j,k}+H\mu)}, \qquad (5.8)
$$

Les deux facteurs restants de l'équation 5.2 sont obtenus en intégrant par rapport à π et θ respectivement.

$$
p(\mathbf{s}|\mathbf{z},\gamma) = \left(\frac{\Gamma(\sum_k\gamma_k)}{\prod_k\Gamma(\gamma_k)}\right)^{T}\prod_{j}\frac{\prod_k\Gamma(n_{j,k}+\gamma_k)}{\Gamma(n_j+\sum_k\gamma_k)}, \qquad (5.9)
$$

$$
p(\mathbf{z}|\alpha) = \left(\frac{\Gamma(\sum_j\alpha_j)}{\prod_j\Gamma(\alpha_j)}\right)^{D}\prod_{d}\frac{\prod_j\Gamma(n_{d,j}+\alpha_j)}{\Gamma(n_d+\sum_j\alpha_j)}, \qquad (5.10)
$$

Distribution a posteriori. La distribution *a posteriori* est obtenue en calculant la probabilité d'observer une thématique et une opinion (variables z et s) connaissant toutes les autres variables. Les notations utilisées ici sont similaires à celles du modèle TS présenté dans le chapitre précédent. Nous utilisons l'exposant "$-p$" pour représenter le sous-ensemble des données d'apprentissage qui exclut le terme à la position p du document courant. L'opinion et la thématique associées à ce terme sont notées s_p et z_p respectivement.

La distribution marginale de s_p et z_p est donnée par la formule suivante :

$$
\begin{aligned}
& p(s_p = k, z_p = j | \mathbf{w}, \mathbf{t}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma) \\
= & \frac{p(z_p = j, s_p = k, w_p, t_p | \mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma)}{p(w_p, t_p | \mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma)} \\
= & \frac{p(\mathbf{w}, \mathbf{t}, \mathbf{s}, \mathbf{z}, \alpha, \beta, \mu, \gamma)}{p(\mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma)} \cdot \frac{p(\mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma)}{p(w_p, \mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma)} \\
= & \frac{p(\mathbf{w}, \mathbf{t}, \mathbf{s}, \mathbf{z} | \alpha, \beta, \gamma, \mu)}{p(\mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p} | \alpha, \beta, \gamma, \mu)} \cdot \frac{1}{w_p} \\
\propto & \frac{p(\mathbf{w}, \mathbf{t}, \mathbf{s}, \mathbf{z} | \alpha, \beta, \gamma, \mu)}{p(\mathbf{w}^{-p}, \mathbf{t}^{-p}, \mathbf{s}^{-p}, \mathbf{z}^{-p} | \alpha, \beta, \gamma, \mu)} \qquad (5.11)
\end{aligned}
$$

Le numérateur et le dénominateur de la fraction ci-dessus sont remplacés par le produit des trois facteurs déjà calculés dans le paragraphe précédent. Cela donne :

$$
\begin{aligned}
& p(s_p = k, z_p = j | \mathbf{w}, \mathbf{t}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma) \\
\propto & \underbrace{\frac{p(\mathbf{w} | \mathbf{s}, \mathbf{z}, \beta)}{p(\mathbf{w}^{-p} | \mathbf{s}^{-p}, \mathbf{z}^{-p}, \beta)}}_{B} \cdot \frac{p(\mathbf{t} | \mathbf{s}, \mathbf{z}, \mu)}{p(\mathbf{t}^{-p} | \mathbf{s}^{-p}, \mathbf{z}^{-p}, \mu)} \cdot \frac{p(\mathbf{s} | \mathbf{z}, \gamma)}{p(\mathbf{s}^{-p} | \mathbf{z}^{-p}, \gamma)} \cdot \frac{p(\mathbf{z} | \alpha)}{p(\mathbf{z}^{-p} | \alpha)} \qquad (5.12)
\end{aligned}
$$

Chacun de ces quatre facteurs peut être simplifié en utilisant la propriété de récursivité de la fonction Γ. Par exemple, le premier facteur, noté B est calculé comme suit :

$$
\begin{aligned}
B &= p(\mathbf{w} | \mathbf{s}, \mathbf{z}, \beta) \cdot p(\mathbf{w}^{-p} | \mathbf{s}^{-p}, \mathbf{z}^{-p}, \beta)^{-1} \\
&= \prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\prod_i \Gamma(n_{i,j,k}^{-p} + \beta + 1)}{\Gamma(n_{j,k}^{-p} + V\beta + 1)} \cdot \left[\prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\prod_i \Gamma(n_{i,j,k}^{-p} + \beta)}{\Gamma(n_{j,k}^{-p} + V\beta)} \right]^{-1} \\
&= \prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\prod_i (n_{w_p,j,k}^{-p} + \beta) \Gamma(n_{i,j,k}^{-p} + \beta)}{(n_{j,k}^{-p} + V\beta) \Gamma(n_{j,k}^{-p} + V\beta)} \cdot \prod_{j=1}^{T} \prod_{k=1}^{S} \frac{\Gamma(n_{j,k}^{-p} + V\beta)}{\prod_i \Gamma(n_{i,j,k}^{-p} + \beta)} \\
&= \frac{n_{w_p,j,k}^{-p} + \beta}{n_{j,k}^{-p} + V\beta} \qquad (5.13)
\end{aligned}
$$

Les trois autres facteurs de la probabilité jointe peuvent être obtenus de la même manière. Au final, la formule de calcul de la probabilité jointe devient :

$$
\begin{aligned}
& p(s_p = k, z_p = j | \mathbf{w}, \mathbf{t}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma) \\
\propto & \frac{n_{w_p,j,k}^{-p} + \beta}{n_{j,k}^{-p} + V\beta} \cdot \frac{n_{j,k,t_p}^{-p} + \mu}{n_{j,k}^{-p} + I\mu} \cdot \frac{n_{j,k}^{-p} + \gamma_k}{n_j^{-p} + \sum_k \gamma_k} \cdot \frac{n_{d,j}^{-p} + \alpha_j}{n_d^{-p} + \sum_j \alpha_j} \qquad (5.14)
\end{aligned}
$$

Enfin, le résultat de cet échantillonnage est utilisé pour estimer les paramètres φ, θ, π et ψ comme suit :

$$\varphi_{j,k,i} = \frac{n_{i,j,k} + \beta}{n_{j,k} + V\beta}, \quad \psi_{j,k,h} = \frac{n_{j,k,h} + \mu}{n_{j,k} + H\mu},$$
$$\pi_{j,k} = \frac{n_{j,k} + \gamma_k}{n_j + \sum_k \gamma_k}, \quad \theta_{d,j} = \frac{n_{d,j} + \alpha_j}{n_d + \sum_j \alpha_j}. \tag{5.15}$$

5.3.3 Intégration de la connaissance a priori

La prise en compte du lexique d'opinion est réalisée de la même manière que pour le modèle TS (voir section 4.4.3).

5.3.4 Régularisation de la modalité "temps"

En examinant le processus génératif du modèle TTS, nous remarquons que différentes étiquettes de temps peuvent être générées pour les termes d'un même document car la génération d'une étiquette de temps se fait au niveau du terme. Or, dans la réalité les termes d'un même document ont tous la même étiquette temporelle, c'est-à-dire celle du document. Cela n'est, en pratique, pas un obstacle pour le déploiement efficace du modèle TTS car celui-ci n'est pas utilisé pour générer des documents. Cependant, comme la modalité "temps" est impliquée dans l'extraction des thématiques, elle peut influencer l'homogénéité de celles-ci. La modalité "temps" est supposée avoir le même poids que la modalité "termes" alors que ce n'est pas le cas (une modalité de temps pour n_d modalités de termes).

Ce problème est classique dans le domaine de la reconnaissance automatique de la parole. Il est généralement résolu en introduisant un paramètre de régularisation qui sert à rééquilibrer les contributions des deux modalités : temps et termes. La même stratégie a également été utilisée dans les modèles TOT [116] et le modèle *Group-Topic model* [117].

Nous adoptons la même stratégie que ces travaux afin de rééquilibrer les contributions des modalités temps et termes du modèle TTS. Le paramètre de régularisation apparaît comme une exponentielle du dernier facteur de l'équation 5.14 ci-dessus. Une valeur naturelle pour ce paramètre est égale à $\frac{1}{n_d}$, ce qui est équivalent à prendre en compte une modalité du temps pour n_d modalités de termes. La distribution *a posteriori* devient donc :

$$\begin{aligned} & p(s_p = k, z_p = j | \mathbf{w}, \mathbf{t}, \mathbf{s}^{-p}, \mathbf{z}^{-p}, \alpha, \beta, \mu, \gamma) \\ \propto \; & \frac{n^{-p}_{w_p,j,k} + \beta}{n^{-p}_{j,k} + V\beta} \cdot \left(\frac{n^{-p}_{j,k,t_p} + \mu}{n^{-p}_{j,k} + I\mu} \right)^{\frac{1}{n_d}} \cdot \frac{n^{-p}_{j,k} + \gamma_k}{n^{-p}_{j} + \sum_k \gamma_k} \cdot \frac{n^{-p}_{d,j} + \alpha_j}{n^{-p}_{d} + \sum_j \alpha_j}. \end{aligned} \tag{5.16}$$

5.3.5 Algorithme d'inférence.

La procédure complète de l'algorithme d'échantillonnage de Gibbs pour le modèle TTS est donné dans l'algorithme 2.

Algorithm 2 Inférence pour le modèle TTS

Require: $\alpha, \beta, \gamma, \mu, T$, un lexique $\mathbb{L}$

Initialiser aléatoirement les matrices Φ, Θ, Π, Ψ.
for iteration $c = 1$ **to** *nbGibbsIterations* **do**
  for document $d = 1$ **to** D **do**
    for $p = 1$ **to** n_d **do**
      Exclure le terme w_p du document d et mettre à jour les variables de comptage
      Tirer une thématique et une opinions pour le terme w_p en utilisant l'équation 5.14
      if le terme w_p est dans le lexique $\mathbb{L}$ **then**
        Changer l'opinion de w_p pour qu'elle corresponde à celle du lexique $\mathbb{L}$
      end if
      Mettre à jour les variables de comptage avec les nouvelles données
    end for
  end for
end for
Mettre à jour les matrices Φ, Θ, Π, Ψ en utilisant les équations 5.15

Complexité algorithmique. L'algorithme d'inférence pour le modèle TTS a la même complexité algorithmique que celui pour le modèle TS. Celle-ci est égale à $O(W \cdot S \cdot T)$ pour chaque itération de l'algorithme (voir section 4.4.2 pour plus de détails).

5.4 Expérimentations

Dans cette section, nous présentons les tests réalisés, la méthodologie d'évaluation ainsi que les résultats obtenus avec le modèle TTS. Afin d'assurer la reproductibilité de ces résultats, nous utilisons comme modèles de comparaison JST et ASUM puisque leurs codes sources sont disponibles. Comme JST et ASUM ne fournissent pas le même type de résultats, nous utilisons un post-traitement spécifique afin d'aligner les résultats de ces deux modèles avec ceux de TTS.

5.4.1 Données et paramètres

Jeux de données. En plus des deux corpus MDSfr et MDSen décrits dans le tableau 4.2, nous utilisons le corpus NYSK (*New York v. Strauss-Kahn*).

NYSK est un corpus d'articles de presse relatifs aux accusations d'agression sexuelle contre l'ancien directeur du FMI [1] Dominique Straus-Kahn en mai 2011 (connue sous l'appellation "affaire DSK"). Nous avons collecté ce corpus en utilisant la plateforme de veille AMIEI (décrite dans le chapitre 6) sur des dizaines de sites de presse anglophone avec la requête : "dsk" OU "strauss-kahn" OU "strauss-khan". Les documents du corpus NYSK sont annotés avec la date de création (du 17/05/2011 au 26/05/2011). Les jeux de données que nous avons créés (MDSfr et NYSK) sont disponibles pour téléchargement sur la plateforme de partage *UCI Repository* [2]. Les codes sources correspondant aux modèles TS et TTS sont également disponibles sur la plateforme Media-Mining [3].

Le tableau 5.1 donne quelques statistiques des trois corpus employés : NYSK, MDSfr et MDSen. La répartition des documents sur le temps est donnée par les graphiques de la figure 5.4.1.

Corpus	Type	Langue	D	V	Annotation	Etiquettes temporelles
MDSfr	critiques	Français	10668	12773	thématique, opinion, temps	mois
MDSen	critiques	Anglais	29379	43834	thématique, opinion, temps	années
NYSK	articles de presse	Français	10421	51188	temps	jours

TABLE 5.1 – Statistiques des corpus MDSfr, MDSen et NYSK.

Paramètres. Les paramètres des modèles TTS, JST et ASUM utilisés pour cette évaluation sont données dans le tableau 5.2. Pour le choix de ces paramètres, nous nous sommes basés sur des expérimentations mais aussi sur les tests effectués avec le modèle TS (cf. section 4.5.1), notamment pour fixer les paramètres α et β. Les expérimentations ont montré que le modèle TTS n'est pas sensible au paramètre μ (*a priori* de la répartition des paires thématiques-opinions dans le temps). Par conséquent, nous avons choisi de fixer ce paramètre à une faible valeur (0.01) afin de ne pas influencer les distributions ψ, c'est-à-dire que ces distributions sont estimées uniquement à partir de données (aucun *a priori*).

1. Fonds Monétaire International
2. http ://archive.ics.uci.edu/ml/
3. http ://mediamining.univ-lyon2.fr/

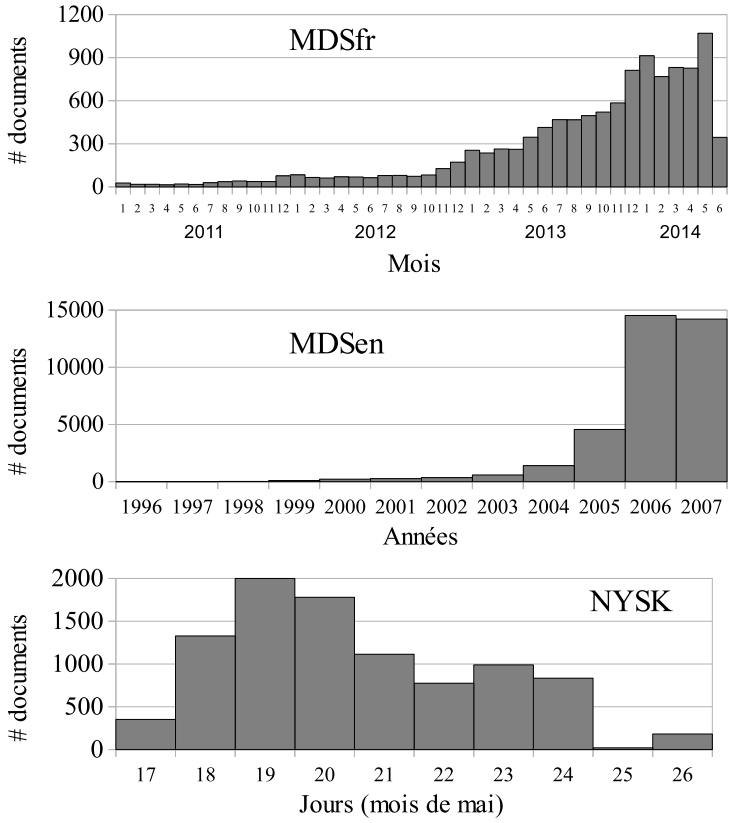


FIGURE 5.4.1 – Répartition des documents sur les étiquettes temporelles.

5.4.2 Méthodologie d'évaluation

Dans cette section, nous décrivons notre approche pour évaluer le modèle TTS ainsi que les deux autres modèles JST et ASUM.

Dans le domaine de l'apprentissage supervisé, les résultats du classement sont confrontés à la vérité terrain afin de calculer l'écart entre la prédiction et la réalité. De manière similaire, notre approche d'évaluation consiste à mesurer le l'écart entre l'estimation et une vérité terrain. Pour réaliser l'évaluation, nous avons choisi de construire la vérité terrain à partir d'un corpus de documents annotés par la thématique, l'opinion et le temps.

Construction de la vérité terrain. La vérité terrain se compose de trois types de distributions de probabilités :

1. Distributions des thématiques sur le vocabulaire (φ)
2. Distributions des thématiques sur les polarités d'opinion (π)
3. Distributions des paires thématiques-opinions sur les étiquettes temporelles (ψ).

Corpus	Paramètres				
	T	α	β	(γ_+, γ_-)	μ
MDSfr	17	2.94	0.06	(1, *)	0.01
MDSen	24	2.08	0.04	(1, *)	0.01
NYSK	24	2.08	0.04	(1, *)	0.01

TABLE 5.2 – Paramètres des modèles utilisés pour l'évaluation. Le symbole "*" désigne une valeur que nous faisons varier dans nos expérimentations.

L'obtention des deux premières distributions se fait de la manière que pour le modèle TS (cf. section 4.5.1). L'obtention de la troisième distribution ψ s'effectue pour une paire thématique-opinion (z, s) et pour une étiquette de temps t de la manière suivante :

1. Affecter chaque document d à la thématique z et la polarité d'opinion s pour lesquelles la probabilité $p(d|s, z)$ est maximale.
2. La probabilité $\psi_{z,s}(t)$ est égale au nombre de documents étiquetés avec t et qui sont affectés à la thématique z et la polarité d'opinion s, divisé par le nombre de documents total D.

Calcul des mesures d'évaluation. Pour évaluer le modèle TTS, nous proposons deux mesures : Q_s et Q_t. La mesure de qualité Q_s correspond à la capacité du modèle à estimer les proportions d'opinions relatives aux thématiques. L'opinion relative à une thématique est exprimée comme une distribution multinomiale. Afin de prendre en compte le caractère flou de cette information, nous définissons la mesure Q_s comme une mesure de distance entre les distributions réelles et les distributions estimées. La deuxième mesure Q_t correspond à la capacité du modèle à estimer la distribution des paires thématiques-opinions dans le temps. La procédure d'évaluation se déroule en deux étapes : mise en correspondance et calcul des mesures (cf. figure 5.4.2).

Etape 1 : mise en correspondance. La mise en correspondance consiste à associer une thématique réelle avec une thématique estimée. Cette étape peut être réalisée manuellement. Cependant, l'évaluation nécessite de répéter cette procédure des dizaines de fois, d'où l'intérêt d'automatiser cette étape. L'association de deux thématiques, réelle et estimée, se fait en se basant sur l'écart entre les distributions de probabilités qui caractérisent ces deux thématiques. Nous proposons de mesurer cet écart par une distance entre ces distributions. Une mesure de distance qui se prête bien aux distributions multinomiales est celle de Kullback-Leibler [4]. Celle-ci est définie pour deux distributions multinomiales A et B comme suit :

$$\mathrm{KLD}(A, B) = \mathrm{KL}(A||B) + \mathrm{KL}(B||A) \tag{5.17}$$

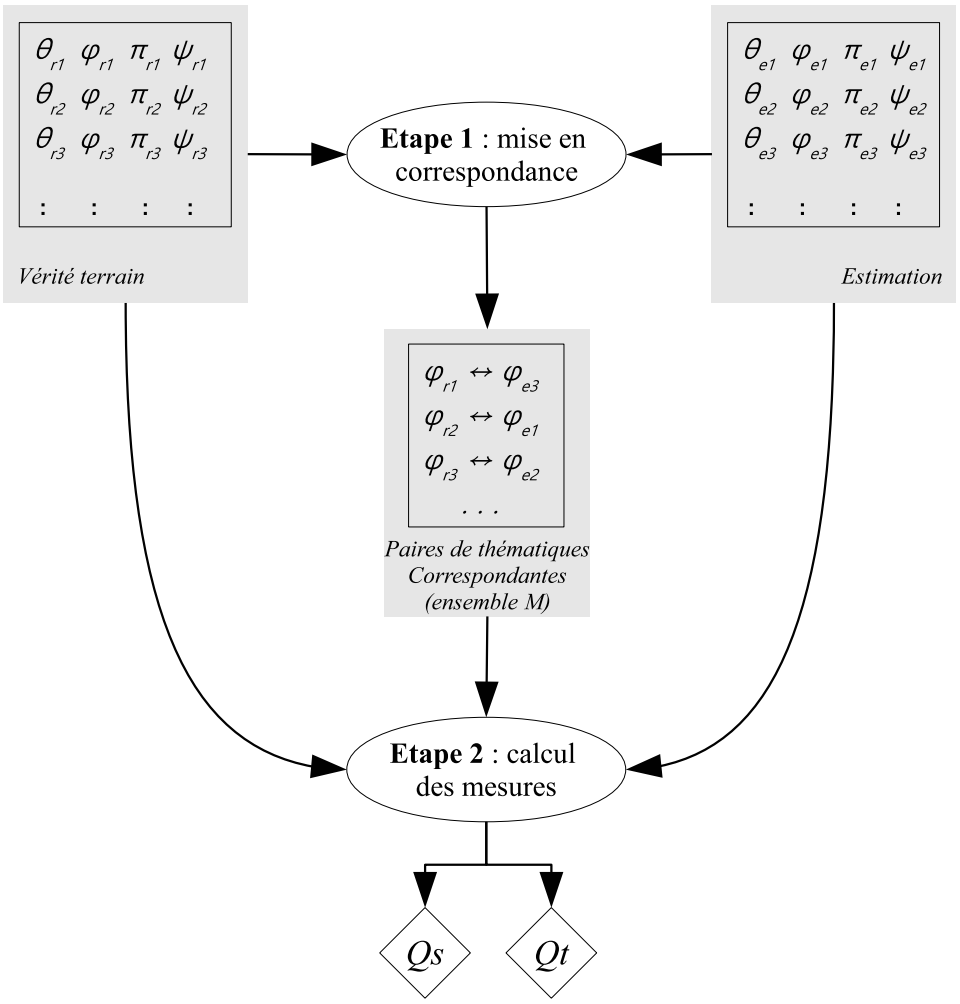


FIGURE 5.4.2 – Méthodologie d'évaluation du modèle TTS.

où $\mathrm{KL}(A||B)$ représente la divergence KL de la distribution A par rapport à la distribution B :

$$\mathrm{KL}(A||B) = \sum_i \left[A(i) \cdot \log \frac{A(i)}{B(i)} \right] \tag{5.18}$$

La mise en correspondance est réalisée en faisant correspondre les deux thématiques pour lesquelles cette distance est minimale et en itérant jusqu'à ce que toutes les paires aient été trouvées.

Pour la simplicité de l'exposé, supposons une modalité d'opinion binaire (positive et négative). Pour l'évaluation du modèle TTS, cette étape de mise en correspondance est réalisée pour les thématiques positives et négatives séparément, i.e., les thématiques réelles positives sont associées aux thématiques estimées positives et *vice versa*.

Soit $\mathbb{L}$, respectivement $\mathbb{E}$, l'ensemble des thématiques réelles, respectivement estimées, définies par leurs distributions sur le vocabulaire. La procédure complète de mise en correspondance pour le modèle TTS est donné par l'Algorithme 3.

Algorithm 3 Mise en corrrespondance des thématiques

Require: Les deux ensembles de thématiques : réelles $\mathbb{L}$ et estimées $\mathbb{E}$
Ensure: $|\mathbb{L}| = |\mathbb{E}|$
1: Initialiser : $\mathbb{M} := \{\}, \mathbb{P} := \mathbb{L} \times \mathbb{E}$.
2: Calculer KLD (r, e) pour chaque paire $(r, e) \in \mathbb{P}$.
3: **for** $z := 1$ **to** $|\mathbb{L}|$ **do**
4: $\quad \mathbb{M} := \mathbb{M} \cup \{(\hat{r}, \hat{e})$ où $(\hat{r}, \hat{e}) = \arg\min_{(r,e)\in\mathbb{P}}$ KLD $(\varphi_r, \varphi_e)\}$.
5: $\quad \mathbb{P} := \mathbb{P} - \{(r, e)$ où $(r = \hat{r} \ \vee \ e = \hat{e})\}$.
6: **end for**
7: **return** $\mathbb{M}$

Il faut noter que la correspondance entre les thématiques de polarités différentes est naturellement fournie par le modèle TTS. Par contre, ce n'est pas le cas des modèles JST et ASUM. Une étape supplémentaire s'avère donc nécessaire afin de faire correspondre les thématiques positives et négatives qui sont similaires. Cette correspondance est réalisée de la même manière en se basant sur la distance KLD.

Etape 2 : calcul des mesures. Soient r une thématique réelle et e une thématique estimée respectivement. Soit $\mathbb{M}$ l'ensemble qui contient le résultat de l'étape précédente (couples de thématiques associées par l'algorithme 3). Chaque couple $(r, e) \in \mathbb{M}$ est caractérisé par une distribution sur les polarités d'opinion. Le calcul de cette distribution est spécifique à chaque modèle. Pour TTS, cette distribution est directement estimée par le modèle (distribution π). Pour JST et ASUM, $p(s|z)$ est obtenue de la même manière que pour les distributions réelles mais avec les annotations estimées. Chaque document d est ré-annoté par l'opinion et la thématique qui maximise la probabilité φ_d.

La première mesure d'évaluation Q_s correspond à la distance entre les distributions qui caractérisent les proportions d'opinions spécifiques aux thématiques (π). Elle est égale à la distance KLD moyenne entre les distributions réelles π_r et estimées π_e correspondants aux couples de thématiques de $\mathbb{M}$:

$$Q_s = \frac{1}{T} \cdot \sum_{(r,e)\in\mathcal{M}} \mathrm{KLD}(\pi_r, \pi_e) \tag{5.19}$$

La deuxième mesure d'évaluation Q_t correspond à la distance entre les distributions qui caractérisent l'évolution temporelle des thématiques et des

opinions (ψ). Elle est égale à la distance entre les distributions réelles ψ_r et estimées ψ_e correspondants aux couples de $\mathbb{M}$:

$$Q_t = \frac{1}{T} \cdot \sum_{(r,e)\in\mathcal{M}} \mathrm{KLD}(\psi_r, \psi_e) \tag{5.20}$$

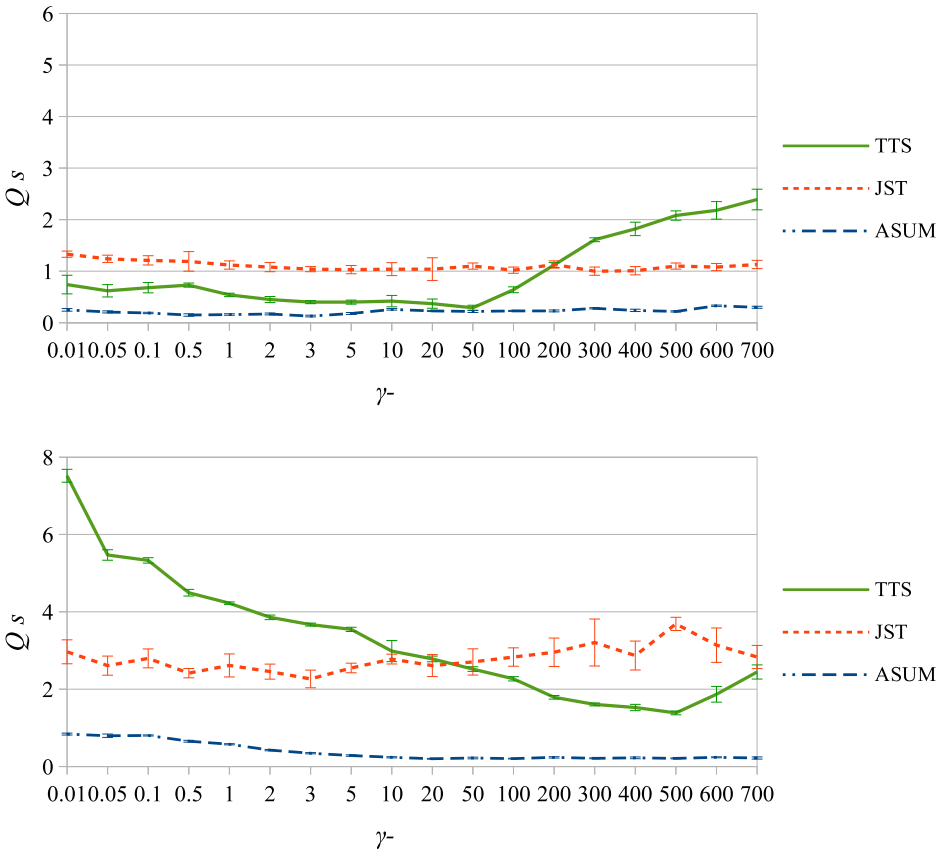


FIGURE 5.4.3 – TTS : Variation de la mesure Q_s (à minimiser) en fonction de γ_- sur les corpus MDSfr (haut) et MDSen (bas). Moyenne et écart-type basés sur 5 initialisations aléatoires.

5.4.3 Résultats

Résultats quantitatifs. Le modèle TTS permet d'abord d'extraire les associations entre les thématiques et les opinions. Pour tester notre modèle sur cette dimension, nous réalisons un test avec différentes valeurs du paramètre γ. En effet, le modèle TTS s'est avéré très sensible à la valeur de ce paramètre. Nous avons choisi de varier ce paramètre et de calculer la mesure de qualité Q_s en fonction. Comme le paramètre γ a la forme d'un vecteur de deux dimensions

(γ_+, γ_-), nous avons choisi de fixer γ_+ à 1 et de faire varier γ_-. En effet, des expérimentations préliminaires ont montré que le modèle n'est pas sensible à la valeur de ces paramètres mais au rapport entre eux $(\frac{\gamma_-}{\gamma_+})$.

La figure 5.4.3 montre la variation de la mesure Q_s en fonction du paramètre γ_- sur les corpus MDSfr et MDSen. Le modèle ASUM obtient le meilleur résultat sur les deux corpus. Sur MDSfr, ASUM atteint $Q_s = 0.13$, suivi de TTS ($Q_s = 0.29$), puis de JST ($Q_s = 1.00$). Sur MDSen, ASUM atteint $Q_s = 0.2$, suivi de TTS ($Q_s = 1.39$), puis de JST ($Q_s = 2.26$). Cette expérimentation montre que ces modèles probabilistes sont plus efficaces quand l'unité thématique est la phrase (ASUM) plutôt que le terme (JST et TTS).

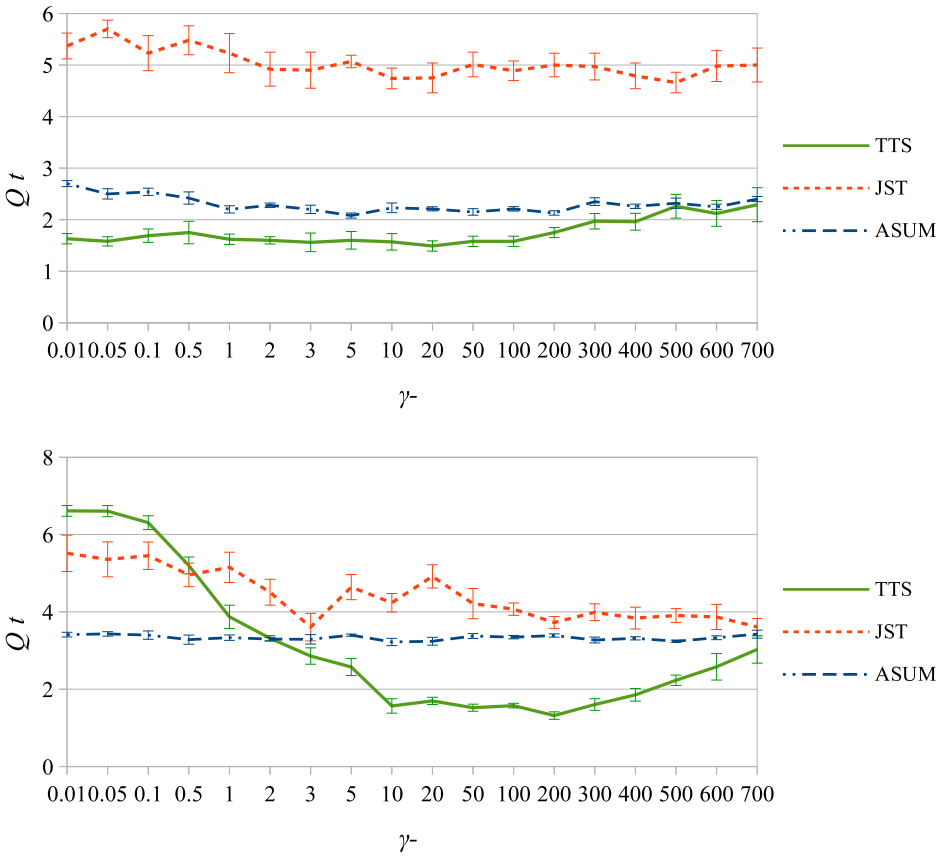


FIGURE 5.4.4 – TTS : Variation de la mesure Q_t (à minimiser) en fonction de γ_- sur les corpus MDSfr (haut et MDSen (bas). Moyenne et écart-type basés sur 5 initialisations aléatoires.

Deux remarques peuvent être faites à partir de ce test :
- Le modèle TTS est sensible au paramètre γ (variation importante de la performance en fonction de γ).
- Le modèle JST est sensible à l'initialisation (écart-types importants par

rapport aux deux autres modèles).

Le deuxième objectif du modèle TTS est l'analyse de l'évolution des associations entre les thématiques et les opinions. Pour tester notre modèle sur cette dimension, nous réalisons la même expérimentation ci-dessus mais nous évaluons les résultats avec la mesure de qualité Q_t qui mesure l'association entre les thématiques, les opinions et le temps. La figure 5.4.4 présente les résultats de cette expérimentation.

Le meilleur résultat en terme de Q_t est atteint par le modèle TTS et ce sur les deux corpus. Sur MDSfr, TTS atteint $Q_t = 1.49$, suivi du modèle ASUM ($Q_t = 2.08$), puis de JST ($Q_t = 4.66$). Sur MDSen, TTS atteint $Q_t = 1.31$, suivi du modèle ASUM ($Q_t = 3.22$), puis de JST ($Q_t = 3.61$). Cela confirme l'utilité d'intégrer l'information temporelle dans le processus de modélisation (comme dans TTS) plutôt que de l'utiliser pour le post-traitement (comme dans ASUM et JST).

Corpus	Q_s	
	Avec le temps	Sans le temps
MDSfr	**0.29**	0.51
MDSen	**1.49**	1.76

TABLE 5.3 – Apport de l'information temporelle à la modélisation des associations thématiques-opinions.

Apport de l'information temporelle à la modélisation des associations thématiques-opinions. Comme cela a déjà été démontré dans [116], l'information temporelle peut aider à mieux séparer les thématiques. Dans ce test, nous proposons d'étudier l'apport de l'information temporelle à la modélisation des associations thématiques-opinions. Le caractère conjoint du modèle TTS devrait aider à extraire des associations plus proches de la réalité.

Nous réalisons deux tests : avec et sans la prise en compte de l'information temporelle. Pour annuler l'effet de l'information temporelle, nous ré-annotons tous les documents avec la même étiquette du temps. Cela rend cette information sans effet sur l'extraction des thématiques et des opinions car la même étiquette temporelle est toujours tirée quelque soient le terme et le document. Le test est réalisé sur les deux corpus MDSfr et MDSen. En se basant sur les tests précédents et afin d'optimiser les résultats, nous avons fixé le paramètre γ_- à 20 pour le corpus MDSfr et 200 pour le corpus MDSen. Tous les autres paramètres restent inchangés (cf. tableau 5.2). Les résultats de ce test sont donnés par le tableau 5.3.

Comme le montrent ces résultats, l'information temporelle contribue positivement à améliorer la modélisation des associations entre les thématiques et les opinions avec le modèle TTS. En effet, l'intégration de cette information

dans le processus de modélisation permet de réduire la distance Q_s entre les distributions des thématiques sur les opinions (réelles et estimées).

Résultats qualitatifs. Dans le tableau 5.4, nous présentons une sélection de thématiques extraites par le modèle TTS à partir des trois corpus MDSfr, MDSen et NYSK. Le constat que nous pouvons faire ici est le même que pour le modèle TS. De manière qualitative, nous remarquons que ces thématiques ont préservé leur sémantique et leur homogénéité, et ce malgré la prise en compte de l'information temporelle. Cependant, il est difficile de valider cette constatation numériquement, i.e., de manière quantitative. Nous nous contentons, dans cette discussion, de présenter ces exemples de ces thématiques.

Par exemple, sur le corpus MDSfr, la thématique z_1 est relative à la vente des chaussures. Nous remarquons l'association des termes positif ("joli" et "qualité") avec la description positive de cette thématique. La description négative a été associée avec le terme "dur". Sur le corpus MDSen, la thématique z_6 à titre d'exemple représente les articles informatiques (ordinateurs et jeux vidéos). Elle a été clairement associée à des termes positives (comme "*good*", "*better*", "*enjoy*") dans sa description positive. De même, cette thématique a été décrite de manière négative avec les termes "*boring*", "*bad*", "*hard*", etc. Enfin, sur le corpus NYSK, de nombreuses thématiques homogènes ont pu être extraites avec le modèle TTS. Par exemple, la thématique z_{10} est relative à l'enquête policière qui a été faite afin de s'assurer de la crédibilité de la plaignante. La thématique z_{12} est relative aux élections présidentielles françaises de 2012 où l'ex-directeur du FMI était considéré comme un candidat potentiel.

De manière générale, la majorité des thématiques extraites avec le modèle TTS sont sémantiquement significatives et peuvent être nommées (manuellement) sans difficulté. Nous aurons l'occasion de fournir des analyses plus fines sur les dimensions de l'opinion et du temps, notamment pour le corpus NYSK, dans le chapitre 6.

5.5 Discussion

Le modèle TTS a pour objectif l'extraction des thématiques et des opinions conjointement à leur évolution temporelle quantitative. Comme cela a déjà été montré, il n'y a aucun modèle de la littérature qui permet une modélisation conjointe de ces trois aspects du texte. Le caractère conjoint du modèle TTS a l'avantage de permettre à l'information temporelle de contribuer à la caractérisation des thématiques et des opinions. Par conséquent, comme l'ont montré les tests réalisés, il permet une meilleure estimation de la distribution des thématiques et des opinions dans le temps.

z_1 : **chaussures**		z_2 : **téléphones**		z_3 : **cuisine**		z_4 : **animalerie**	
positif	*négatif*	*positif*	*négatif*	*positif*	*négatif*	*positif*	*négatif*
couleur	lavage	pratique	brancher	famille	pouvoir	vraiment	image
texture	taille	lecteur	*compliqué*	électronique	échapper	air	image
chaussures	semelles	bon	connexion	siemens	chauffer	croquettes	céréales
joli	*dur*	samsung	*impossible*	friteuse	ondes	chat	produit
cuir	chaussures	top	boîtier	retourner	rapidement	sucre	livrer
semelle	fabrication	autonomie	batterie	qualité	bouilloire	aimer	gamelle
pieds	supporter	tester	allumer	minutes	dessous	liesse	*déçu*
redire	acheter	esthétique	*bug*	démonter	durée	pastilles	prix
porter	double	écran	seul	recommander	capacité	adopter	grand
qualité	bout	comparer	photo	appareil	*mauvais*	dessin	description

z_5 : **video**		z_6 : **computer and video games**		z_7 : **beauty**		z_8 : **software**	
positif	*négatif*	*positif*	*négatif*	*positif*	*négatif*	*positif*	*négatif*
movie	film	game	way	hair	smell	use	computer
story	character	play	player	product	scent	software	xp
scene	actor	one	*boring*	great	shaver	work	upgrade
good	movie	fun	*puzzle*	shave	*thick*	support	internet
best	*horror*	good	run	dry	eye	file	crash
actor	role	graphic	*long*	feel	*iron*	system	manual
star	war	level	*bad*	recommend	*disappointed*	feature	*slow*
performance	*bad*	better	*hard*	clean	*irritate*	easy	connect
director	dvd	enjoy	fight	shampoo	blow	help	laptop
casting	expect	gameplay	screen	cream	trimmer	update	technical

z_9 : **allegation**		z_{10} : **investigation**		z_{11} : **Christine Lagarde**		z_{12} : **French elections**	
positif	*négatif*	*positif*	*négatif*	*positif*	*négatif*	*positif*	*négatif*
allegation	*accused*	evidence	strauss	lagarde	strauss	french	strauss
strauss	*guilty*	told	kahn	minister	next	political	sarkozy
kahn	*criminal*	time	hotel	imf	lagarde	socialists	president
charge	*deny*	court	dna	candidate	nation	party	france
victim	media	*assault*	new	french	crisis	leader	*assault*
arrest	french	investigation	bail	economy	european	dsk	party
police	kahn	consensus	investigation	support	possible	newspaper	image
sexual	hotel	believe	maid	lead	member	candidate	aubry
manhattan	chief	brafman	case	debt	unit	journalist	election
defend	riker	staff	lawyer	develop	first	campaign	french

TABLE 5.4 – Termes de probabilités élevées pour une sélection de thématiques extraites avec le modèle TTS sur les corpus MDSfr (haut), MDSen (milieu) et NYSK (bas). Les termes du lexique sont coloriés en vert/souligné (positifs) ou en rouge/italique (négatifs).

Le modèle TTS peut être utilisé pour plusieurs applications, comme l'analyse de tendances, la e-réputation et de manière plus générale l'analyse exploratoire des données textuelles sur les trois axes : thématiques, opinions et temps.

Par ailleurs, nous avons proposé une nouvelle méthodologie pour l'évaluation du modèle TTS. Comme cela a déjà été discuté, les méthodes de la littérature se sont avérées insuffisantes pour évaluer le modèle selon les trois dimensions (thématiques, opinions, temps). Les deux mesures que nous avons proposées Q_s et Q_t ont permis de tenir compte du caractère multidimensionnel du modèle TTS à travers l'étape de mise en correspondance des thématiques et la quantification de l'écart entre l'estimation que nous faisons et une vérité terrain.

Cependant, en examinant le déroulement de l'évaluation, notamment l'étape de mise en correspondance des thématiques, nous pouvons proposer un axe d'amélioration de ce travail. En effet, dans cette évaluation, chacune des thématiques réelles est obligatoirement mise en correspondance avec une thématique estimée. Or, comme l'extraction des thématiques est complètement non supervisée, des thématiques nouvelles et inexistantes parmi les thématiques réelles, peuvent être extraites. Par exemple, sur le corpus des critiques MDSfr, une thématique relative aux "envois de colis" a été extraite par le modèle TTS. Celle-ci ne correspond à aucune catégorie de produits vendus sur le site Amazon et par conséquent à aucune thématique réelle. Cependant, pour l'évaluation, elle doit être mise en correspondance avec une thématique réelle. Notons enfin que cette situation est marginale et ne concerne qu'une minorité de thématiques dans le cas de nos expérimentations. Une analyse manuelle des résultats de cette étape a montré que la plupart des thématiques mises en correspondance sont sémantiquement liées et correspondent clairement à la même catégorie de produits.

Le modèle TTS étant basé sur la théorie des réseaux bayésiens est extensible afin d'inclure d'autres axes d'analyse. Une des pistes intéressantes serait d'inclure l'information structurelle des données [87, 97]. Certains types de données, comme les données issues des réseaux sociaux, ont une structure (utilisateurs, liens, ..). Plusieurs travaux sont proposés pour l'exploitation de cette structure, comme par exemple la détection de communautés (groupes d'utilisateurs interconnectés). La combinaison de cette connaissance avec nos travaux permettrait de répondre à des questions du genre : les membres d'une même communauté, parlent-ils des mêmes sujets ? ont-ils des opinions similaires ? comment le changement d'opinion au niveau de certains utilisateurs peut-il se propager au sein de la communauté ? et dans quelle mesure ?

Chapitre 6

Implémentation

Résumé. *Dans ce chapitre, nous présentons le cadre industriel de la thèse : l'entreprise AMI Software et la plateforme de veille AMIEI au sein de laquelle nos travaux ont été intégrés. Ça sera aussi l'occasion de présenter nos outils de visualisation des résultats pour les deux contributions : l'analyse d'opinions avec la méthode hybride ainsi que l'analyse conjointe et dynamique des thématiques et opinions avec le modèle TTS. Ces travaux sont intégrés sous forme de deux composants au sein de la plateforme AMIEI. A la fin de ce chapitre, nous présentons des cas d'étude sur des sujets qui ont fait l'actualité pendant cette thèse.*

Sommaire

6.1 Introduction

La veille sur le Web consiste en une surveillance automatisée d'une ou plusieurs sources d'informations sur le Web. On distingue généralement plusieurs types de veille sur le Web : veille stratégique, concurrentielle, économique, documentaire, scientifique, etc. Dans plusieurs domaines, la veille est un service incontournable sur lequel se base la prise de décisions importantes et stratégiques. Voici quelques exemples d'utilisation de la veille sur le Web :

- Veille économique et concurrentielle : permet aux entreprises d'être à l'écoute de leurs clients et de rester informées sur leur environnement économique (fournisseurs, concurrents, ...).
- E-réputation : la e-réputation est définie comme l'opinion commune véhiculée par le Web et perçue par les internautes par rapport à une entité (entreprise, marque, homme politique, ...). La veille sur le Web permet de déceler et de traquer ces opinions dans l'objectif d'une meilleure gestion de l'image sur le Web.

Les travaux réalisés dans le cadre de cette thèse s'insèrent dans un contexte général de veille sur le Web. Cette thèse s'est déroulée dans l'entreprise AMI Software [1]. Créée au début des années 2000, AMI Software (raison sociale : Go Albert), est une entreprise de droit français, initialement spécialisée dans les moteurs de recherche. Aujourd'hui, AMI Software développe une multitude d'outils de veille et d'intelligence économique, notamment la plateforme de veille AMIEI (*AMI Enterprise Intelligence*) qui permet la mise en œuvre des différentes phases du processus de veille sur le Web.

Dans ce chapitre, nous présentons brièvement l'intégration de nos travaux au sein de la plateforme AMIEI. Nous montrons principalement l'intégration de deux travaux sous forme de composants : l'analyse d'opinions avec la méthode hybride présentée dans le chapitre 3 et la modélisation dynamique des thématiques et des opinions avec le modèle TTS présenté dans le chapitre 5. Ce sera aussi l'occasion de présenter les méthodes de visualisation de résultats que nous avons adoptées et dont une partie a été publiée dans [19].

Nous commençons par présenter la plateforme de veille AMIEI (section 6.2). Ensuite, nous présentons les deux composants : AMI-Sent (section 6.3) et AMI-Trend (section 6.4). Enfin, nous présentons deux cas d'étude sur des données issues de l'actualité collectées avec la plateforme AMIEI au cours de la thèse (section 6.5).

1. http ://www.amisw.com/

6.2 La plateforme de veille AMIEI

La plateforme AMIEI est une solution logicielle destinée à répondre à l'ensemble du processus de veille des entreprises dans des contextes divers tels que l'intelligence économique, la veille technologique, ou l'e-réputation [103]. La plateforme AMIEI consiste en un ensemble de modules indépendants qui permettent de mettre en œuvre les quatre principales phases d'un processus de veille (voir figure 6.2.1), à savoir, l'acquisition de l'information, la capitalisation, l'analyse et enfin le partage et la diffusion.

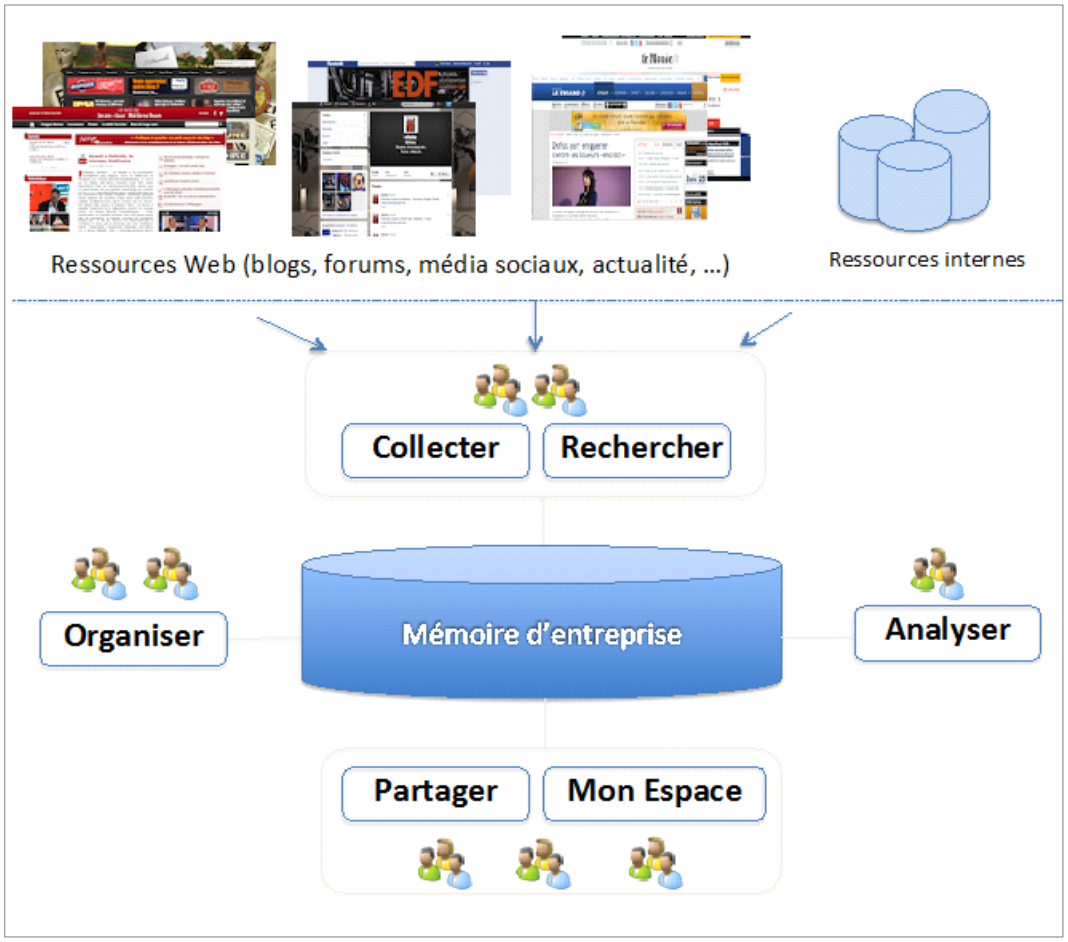


FIGURE 6.2.1 – Processus général de veille au sein de la plateforme AMIEI.

6.2.1 Acquisition

L'acquisition de l'information peut se faire soit avec :
- Un moteur de recherche permettant de faire des recherches ponctuelles.
- Un automate de collecte pour des opérations récurrentes de suivi.

- Un module permettant la contribution des utilisateurs du logiciel (entrée manuelle d'information)
- Un module pour collecter et capitaliser des informations à la volée, en surfant sur le web via un plug-in du navigateur Web.

6.2.2 Capitalisation

La plateforme AMIEI est construite autour d'une base de données pour capitaliser sous une forme organisée et maîtrisée les documents qui ont été collectés. Cette base de données, appelée "mémoire d'entreprise" constitue un capital important de données. Au fur et à mesure de l'exploitation du logiciel, elle favoris le croisement d'informations entre elles, permet de retrouver des connaissances enregistrées depuis plusieurs mois ou années, pour devenir un véritable lieu de partage de connaissances et d'information. Cette mémoire d'entreprise est structurée sous forme d'un plan de classement (hiérarchie de catégories) qui reçoit l'ensemble des résultats de la phase d'acquisition et sur laquelle s'appuient les fonctionnalités d'analyse et de partage.

6.2.3 Analyse

L'objectif de cette phase est l'exploration des données acquises, en vue d'en extraire une information utile et pertinente. Divers outils d'analyse statistique, de fouille de textes et de visualisation sont proposés. Ils permettent de fournir une cartographie des données collectées selon différents critères (temps, sources, etc.) et permettent la mise en évidence de tendances, la détection d'informations de rupture (signaux faibles) ou l'extraction automatique d'entités nommées (personne, lieu, organisation, concepts généraux, etc.).

6.2.4 Partage

Le partage et la diffusion des informations acquises et validées, ainsi que les résultats de l'analyse, se sont à travers deux points d'accès principaux :

- Un portail de consultation : un module permettant la recherche et le partage des informations organisées par thème avec une gestion des droits d'accès à partir de profils prédéfinis. Le portail qui dispose d'un moteur de recherche intégré, permet une navigation dans les résultats de collecte. Il permet aussi aux utilisateurs de s'abonner à des fils RSS afin d'accéder aux contenus de la plateforme à partir d'applications tierces.
- Mon espace : un module permettant de personnaliser, pour chaque utilisateur, son accès à la plateforme AMIEI. Ainsi, un veilleur peut suivre son activité de veille, les documents collectés, les volumes d'information produits par chaque source, les statistiques d'utilisation, etc.

6.3 Contribution 1 : le composant AMI-Sent

Notre méthode hybride pour l'analyse d'opinions, décrite dans le chapitre 3, a été intégrée au sein de la plateforme AMIEI sous la forme d'un composant (module indépendant qui interagit avec les autres modules). Le composant AMI-Sent repose sur le module de collecte pour l'acquisition des données directement à partir d'un plan de veille. Ensuite, plusieurs prétraitements sont appliquées sur ces données (segmentation, nettoyage, représentation, etc.) avant d'y appliquer la méthode d'analyse d'opinions. Enfin, les résultats de l'analyse sont restitués via une visualisation adaptée.

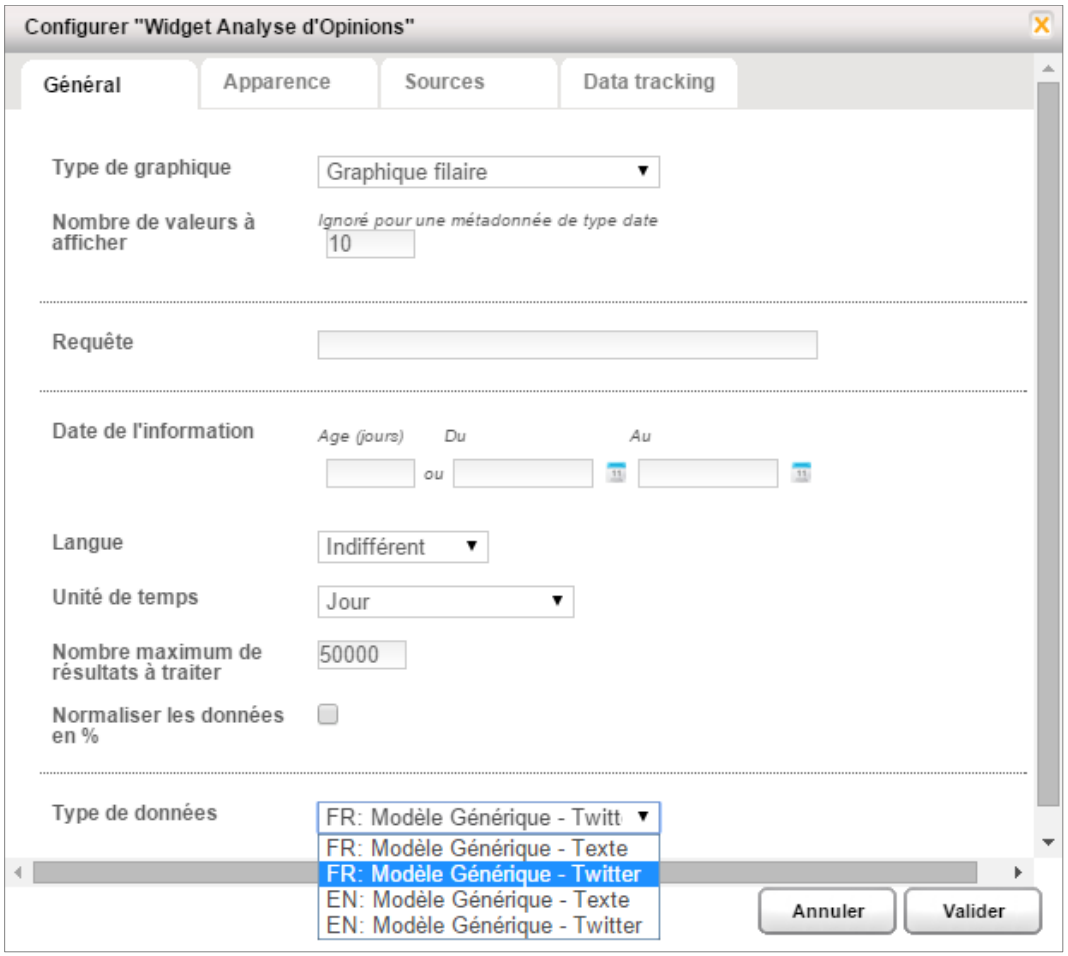


FIGURE 6.3.1 – Composant AMI-Sent : interface de configuration.

6.3.1 Configuration du composant AMI-Sent

Pour la mise en œuvre de la méthode hybride, rappelons qu'il faut disposer d'un corpus d'apprentissage annoté. Afin de faciliter l'utilisation de ce compo-

sant, nous avons construit quatre modèles d'analyse d'opinions sur la base de quatre corpus différents. L'utilisateur choisit le modèle qui correspond à ses besoins selon deux critères : le type et la langue des données à analyser. Les corpus et les prétraitements utilisés pour la construction de ces quatre modèles sont donnés dans le tableau 6.1.

Modèle	Corpus	Prétraitements
FR-TEXT	critiques de films, hôtels, restaurants [112]	mots vides, valeurs numériques
FR-TW	4 783 tweets collectés et annotés manuellement	mots vides, valeurs numériques, mots clés Twitter (RT, via, ..)
EN-TEXT	critiques MDSen [10]	mots vides, valeurs numériques
EN-TW	tweets SemEval [122]	mots vides, valeurs numériques, mots clés Twitter (RT, via, ..)

TABLE 6.1 – Corpus et prétraitements utilisés pour la construction des modèles d'analyse d'opinions.

La configuration du composant AMI-Sent est définie par la langue et le type de données. Le composant AMI-Sent fonctionne selon quatre configurations différents correspondant à la combinaison des deux langues Français et Anglais avec les deux types de données Générique et Twitter. Le type de données Générique correspond à n'importe quel type de documents, mis à part les tweets auxquels nous avons réservé un prétraitement particulier pour leurs spécificités. Les quatre configurations sont donc "Français-Twitter", "Français-Générique", "Anglais-Twitter" et "Anglais-Générique" (cf. figure 6.3.1). Nous avons créé les modèles d'analyse correspondant à ces quatre configurations prédéfinis mais le composant est extensible à d'autres configurations.

6.3.2 Visualisation des résultats

Les résultats de l'analyse d'opinions sont présentées selon plusieurs dimensions, comme le montre la figure 6.3.2.

Répartition des documents sur les polarités d'opinion. La première visualisation montre la répartition des documents analysés sur les polarités de l'opinion (cf. figure 6.3.3). Cette information est restituée à l'aide d'un graphique en secteurs où chaque secteur correspond à une polarité d'opinion.

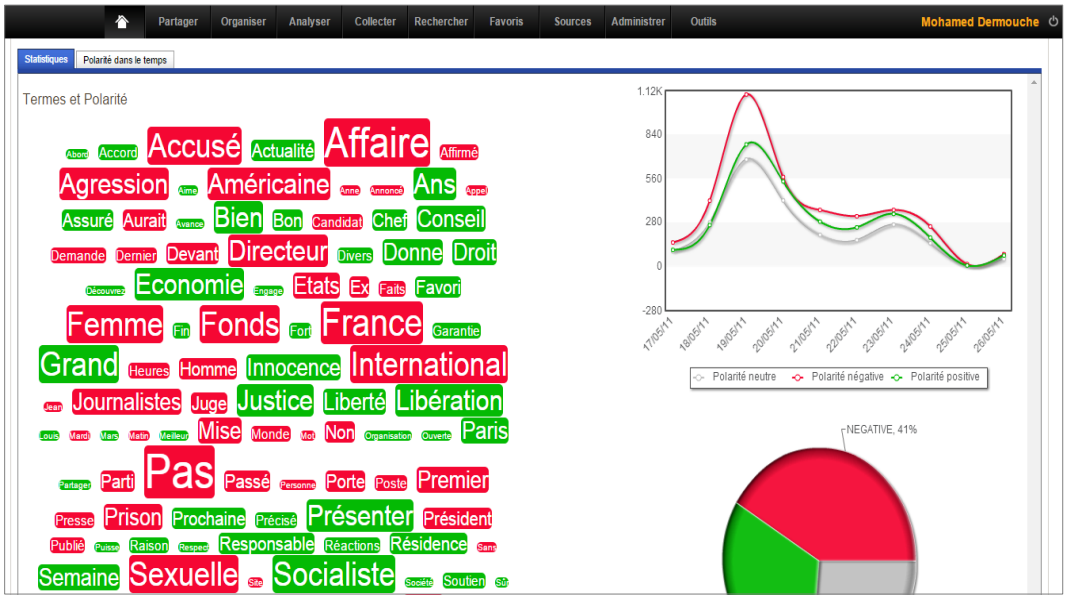


FIGURE 6.3.2 – Composant AMI-Sent : vue d'ensemble de la visualisation des résultats.

Evolution volumétrique. En se basant sur l'information temporelle (étiquettes de temps associées aux documents), le résultat de l'analyse est projeté sur l'axe temporel afin de visualiser l'évolution volumétrique des polarités (figure 6.3.4).

Nuage de termes. Les résultats de l'analyse d'opinions sont également visualisés par un nuage de termes. Pour cela, nous nous appuyons sur les résultats retournés par la méthode hybride et nous calculons un score de confiance pour chaque document classé. Le score est compris entre 0 et 1 et il est calculé, pour un document d, de la manière suivante :

- Les probabilités $p(c_i|d)$ sont triées telles que : $p(c_m|d) > p(c_n|d) > ... > p(c_p|d)$, où c_i sont les classes d'opinion (polarités).
- Le score de confiance $\text{Confiance}(d) = p(c_m|d) - p(c_n|d)$. Il représente l'écart entre la classe la plus probable et la deuxième classe la plus probable. Plus cet écart est important, plus l'association entre le document d et la classe d'opinion m est forte.

Ensuite, nous proposons d'expliquer l'affectation d'un document d à une classe d'opinion par les termes qui ont le plus contribué à cette affectation. Ceci est réalisé de la manière suivante :

- Soit c la classe d'opinion du document d (classe la plus probable).
- Evaluer chaque terme w_i du document d selon un critère de spécificité (pouvoir discriminatif du terme au regard de la classe d'opinion). Ici, nous choisissons comme critère le gain informationnel (IG). Ensuite, trier les

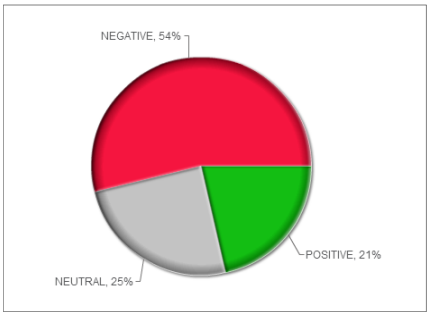


FIGURE 6.3.3 – Composant AMI-Sent : répartition des documents sur les polarités d'opinion.

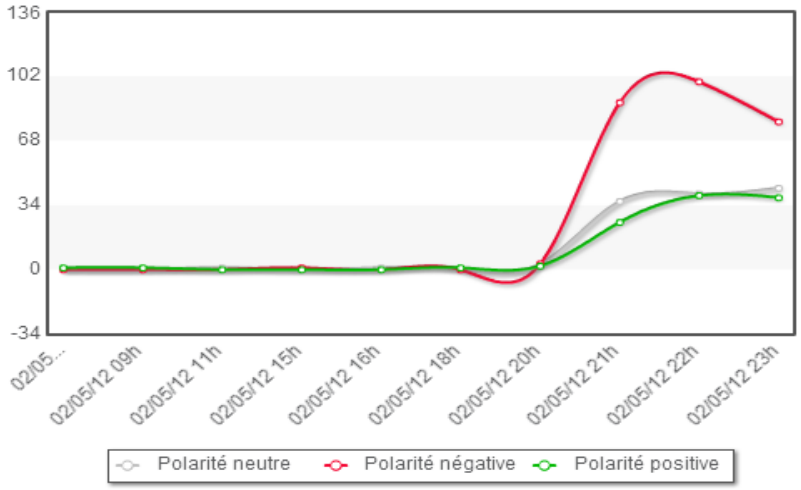


FIGURE 6.3.4 – Composant AMI-Sent : évolution volumétrique de chaque polarité d'opinion.

termes w_i du document selon ce critère : $\text{IG}(w_h|c) > \text{IG}(w_i|c) > ... > \text{IG}(w_j|c)$.

– Les K premiers termes sont ceux qui "expliquent" le mieux cette affectation.

Nous précisons que les termes discriminants de deux classes différentes sont deux ensembles disjoints. En effet, un terme ne peut être responsable de classer un texte que dans une seule classe.

Enfin, les termes discriminants issus de tous les documents du corpus sont fusionnés afin de générer le nuage de termes. Celui-ci prend en compte deux caractéristiques des termes : la fréquence, en la rapportant sur la taille graphique du terme et la polarité en la rapportant sur sa couleur. Pour des raisons d'ergonomie et de lisibilité, la liste des termes est filtrée en prenant un petit ensemble K (en général quelques dizaines) parmi les plus fréquents.

Nuage de termes évolutif. Nous exploitons l'information temporelle des documents (la date de création ou la date de collecte) afin de projeter le nuage de termes sur la dimension temporelle. Cela fournit une visualisation plus fine des termes discriminants et montre leur évolution dans le temps. Afin d'optimiser le rendement de cette visualisation, nous nous sommes appuyés sur la technique de visualisation par *fisheye* (cf. figure 6.3.5).

FIGURE 6.3.5 – Composant AMI-Sent : visualisation basée sur la technique *fisheye*.

6.4 Contribution 2 : le composant AMI-Trend

Le modèle TTS a été implémenté au sein de la plateforme AMIEI sous forme d'un composant indépendant qui interagit avec les modules de la plateforme. Le composant AMI-Trend permet de récupérer les résultats de la collecte afin de les analyser. Ensuite, ce composant permet d'implémenter le modèle TTS présenté dans le chapitre 5.

6.4.1 Modes d'utilisation du composant AMI-Trend

Le composant AMI-Trend offre beaucoup de souplesse dans son utilisation. En se basant sur la même implémentation, celui-ci peut être utilisé pour quatre objectifs différents :

1. **Extraire les thématiques :** le composant AMI-Trend peut être facilement utilisé pour l'extraction de thématiques de manière identique au

modèle LDA. En effet, le fait de considérer une seule polarité d'opinion et une seule étiquette de temps équivaut à annuler ces deux dimensions et retrouver un modèle équivalent à LDA. Dans des cas d'utilisation où l'opinion et le temps ne constituent pas un axe d'intérêt, cette configuration peut devenir utile car elle permet de réduire le temps d'exécution.

2. **Extraire les thématiques et les opinions relatives :** le fait de considérer la même étiquette temporelle pour tous les documents à analyser revient à annuler cette dimension. Cette configuration peut être utilisée au cas où l'information temporelle n'apporte pas ou peu de contribution à la modélisation, par exemple, si les documents sont collectés sur une période de temps très limitée.
3. **Modéliser l'évolution des thématiques :** dans le cas où l'on est intéressé par l'évolution des thématiques indépendamment des opinions, il est possible d'utiliser cette configuration du composant AMI-Trend. Cela revient à considérer une seule polarité d'opinion et de rendre le modèle ainsi équivalent au modèle TOT présenté dans la section 5.2.2. La seule différence concerne la nature de la variable temporelle (continue dans TOT et discrète dans TTS).
4. **Modéliser l'évolution des thématiques et des opinions relatives :** le dernier mode de fonctionnement du composant AMI-Trend correspond à la configuration où les trois dimensions du texte sont considérés (thématiques, opinions et temps). Ce mode permet d'extraire les thématiques, les opinions relatives à ces thématiques et la dynamiques des thématiques et des opinions.

6.4.2 Visualisation des résultats

Les résultats de l'analyse conjointe et/ou dynamique avec le composant AMI-Trend sont visualisés selon trois dimensions : thématiques, associations thématiques-opinions et évolution temporelle des thématiques et des opinions. Dans cette section, nous décrivons très brièvement ces techniques de visualisation. Nous aurons l'occasion de les illustrer dans la section suivante (études de cas).

Description des thématiques. Afin de décrire les thématiques, nous avons choisi deux types de visualisation : les termes probables et les documents représentatifs. Les termes probables sont obtenus par les distributions caractérisant les thématiques (distributions φ). Ainsi, les K termes les plus probables sont sélectionnés et représentés en nuage de termes proportionnellement à leurs probabilités.

Les termes probables offrent une visualisation compacte et efficace des thématiques. Cependant, les termes ne révèlent pas tout sur le contenu des

thématiques à cause de l'insuffisance du contexte. Nous avons choisi de compléter cette visualisation par l'affichage d'un ensemble de documents représentatifs associés à chaque thématique extraite. Les documents représentatifs sont choisis parmi les documents qui ont de fortes probabilités d'appartenance à une thématique en se basant sur les distributions caractérisant les documents (distributions θ). La visualisation se fait en sélectionnant les M documents les plus représentatifs pour une thématique et les représenter par leurs titres (quand le titre est une information disponible) ou à défaut par un résumé. Pour l'obtention de ces résumés, nous faisons appel à des outils de résumé automatique fournis par la plateforme AMIEI.

Opinions relatives aux thématiques. Pour chaque thématique extraite avec le modèle TTS, il existe une distribution multinomiale sur les polarités d'opinion (distributions π). Nous visualisons cette information à l'aide d'un graphique en secteurs, avec des codes couleurs correspondant aux polarités de l'opinion (vert pour le positif, rouge pour le négatif et gris pour le neutre).

Evolution des thématiques et des opinions. L'information concernant la dynamique des thématiques et des opinions relatives est restituée à l'utilisateur via un graphique en courbes. Les courbes correspondent aux différentes polarités de l'opinion associées à la thématique. Les valeurs des ordonnées correspondent aux quantités de données associées à une thématique et à une opinion à un temps t. Cette quantité est calculée par le nombre de documents associées à la thématique et l'opinion au temps t en maximisant la probabilité. Afin d'obtenir cette information, les distributions ψ produites par le modèle TTS sont converties en termes de nombres de documents. Le nombre de documents associés avec une thématique z, une opinion s et une étiquette temporelle t, noté $\text{nbDocs}_{z,s}(t)$, est calculé comme suit :

$$\text{nbDocs}_{z,s}(t) = \psi_{z,s,t} \cdot \pi_{z,s} \cdot \text{topicSize}(z) \tag{6.1}$$

Où $\text{topicSize}(z)$ est le nombre de documents associés à la thématique z en maximisant la probabilité $\theta_{d,z}$.

6.5 Etudes de cas

Afin de montrer l'intérêt de nos contributions, nous avons travaillé sur deux études de cas : le débat présidentiel relatif aux élections françaises (2012) et l'affaire DSK (2011). Pour le débat présidentiel, il s'agit d'analyser les polarités d'opinion avec la méthode hybride. Pour l'affaire DSK, il s'agit d'extraire les thématiques traitées dans la presse et les prises de positions par rapport à ces thématiques (proportions d'opinions) ainsi que leur évolution temporelle.

6.5.1 Le débat présidentiel (mai 2012)

FIGURE 6.5.1 – Nuage de termes obtenu à partir du corpus de tweets liés au débat présidentiel.

Le corpus d'étude est composé de 50 000 tweets collectés par la plateforme de veille AMIEI dans la soirée du 02 Mai 2012 en faisant une restriction sur le tag "#ledebat". Ces tweets sont relatifs au débat télévisé du second tour de l'élection présidentielle française de 2012 ayant opposé F. Hollande et N. Sarkozy. Nous appliquons comme prétraitement la suppression de mots outils et des valeurs numériques.

Les figures 6.5.1 et 6.5.2 représentent la visualisation du résultat d'analyse du corpus Politique. Pour une meilleure lisibilité, seulement une sélection de termes fréquents est représentée ici. La polarité des termes est représentée par une couleur (vert pour le positif et rouge pour le négatif). Ces résultats nous ont permis de cerner les termes et les concepts les plus importants dans chaque catégorie d'opinion. A partir de ces résultats, nous pouvons tirer plusieurs enseignements. En voici quelques uns :

- Le concept de "changement" (slogan phare de la campagne du candidat F. Hollande) est largement repris par les internautes, et ce de manière positive.
- "melenchon2012" (compte officiel de compagne du candidat J.-L. Mélenchon) est souvent associé à des opinions négatives. En effet, ce compte est

FIGURE 6.5.2 – Evolution du nuage de termes obtenu à partir du corpus de tweets liés au débat présidentiel.

particulièrement critique envers les deux candidats (notamment N. Sarkozy) et n'a cessé de twitter des critiques acerbes tout au long du débat.
- "DSK" est marqué négativement. En effet, beaucoup d'utilisateurs (notamment des soutiens de F. Hollande) ont commenté négativement la référence de N. Sarkozy à l'affaire DSK durant le débat, et l'ont interprété comme un manque d'arguments sur d'autres sujets plus sérieux.
- Le concept "Sarko" est marqué négativement. En effet, cette abréviation du nom du candidat N. Sarkozy est surtout utilisée par ses détracteurs et non par ses soutiens.

6.5.2 L'affaire DSK (mai 2011)

Le corpus NYSK présenté dans la section 5.4.1 est une collection d'articles de presse relatifs aux accusations d'agression sexuelle contre l'ancien directeur du FMI Dominique Strauss-Kahn (affaire DSK). Cette affaire a suscité beaucoup d'intérêt en France et à travers le monde avec parfois des polémiques et des controverses [2]. Entre le scandale familial, la théorie du complot, les élections présidentielles françaises et les conséquences politiques, l'histoire a été abordée de tous cotés et a pris beaucoup de tournants.

Dans cette étude, nous proposons d'analyser ce corpus avec le modèle TTS pour l'extraction de la dynamique et des associations entre les thématiques et les opinions. Les paramètres du modèle TTS sont fixés aux valeurs par défaut

2. http ://fr.wikipedia.org/wiki/Affaire_Dominique_Strauss-Kahn/

($\alpha = \frac{50}{T}$ et $\beta = \frac{1}{T}$). Nous utilisons la méthode de mise à jour dynamique proposée dans la section 4.5.4 afin de fixer le paramètre γ. Les résultats obtenus sont représentés par une sélection de thématiques extraites. Pour chacune de ces thématiques, nous visualisons :

- Le nuage de termes pour les deux polarités d'opinion (positive et négative)
- L'évolution temporelle de la thématique conjointement avec les deux polarités d'opinion en terme de nombre de documents. Sur l'axe des ordonnées, nous représentons la quantité de données (exprimée en nombre de documents) liée à la thématique et à l'opinion en même temps. La méthode de calcul du nombre de documents a déjà été donnée dans la section 6.4.1.

Nous présentons les résultats pour une sélection de six thématiques dans les figures 6.5.3-6.5.8.

Sur la base de cette analyse, il est possible de tirer quelques renseignements en les liant aux événements de l'actualité. En voici quelques uns :

- La thématique z_9 relative aux premières accusations contre DSK se répartit majoritairement sur les 4 premiers jours suivant le début de l'affaire. La thématique est fortement associée à une opinion positive. Cela reflète le fait que les médias prenaient cette affaire avec beaucoup de précaution durant les premiers jours. Nous remarquons aussi sur le tableau 5.4 l'utilisation de quelques mots clés qui résument la réaction immédiate des médias comme "*allegation*", "*victim*", "*accused*", etc. Cela reflète le soutien et le choc des médias durant cette brève période.
- La thématique z_{10} est liée à l'enquête menée par la NYPD (*New York City Police Department*) afin de s'assurer de la crédibilité de la plaignante. Cette thématique présente deux pics d'apparition correspondant aux dates du 19 mai et 24 mai. Ces deux dates correspondent respectivement à la date à laquelle l'enquête a été engagée et la date de parution des premiers résultats de l'enquête, notamment les résultats du test ADN qui étaient positifs.
- La thématique z_{12} est relative aux élections présidentielles françaises de 2012 pour lesquelles DSK était considéré comme un candidat potentiel. L'opinion relative à cette thématique est mitigée. Ce résultat peut être lié à un sondage réalisé par l'institut CSA [3] quelques jours après le début de l'affaire. Ce sondage faisait état d'une opinion publique mitigée par rapport à la candidature de DSK et plus généralement par rapport à la popularité du parti socialiste. Le sondage conclut que cette affaire n'était pas si catastrophique pour le parti socialiste et que la victoire des socialistes était encore possible même sans DSK et la suite des événements leur a donné raison.

3. http ://www.csa.eu/multimedia/data/sondages/data2011/opi20110516-les-premieres-consequences-politiques-de-l-affaire-dsk.pdf

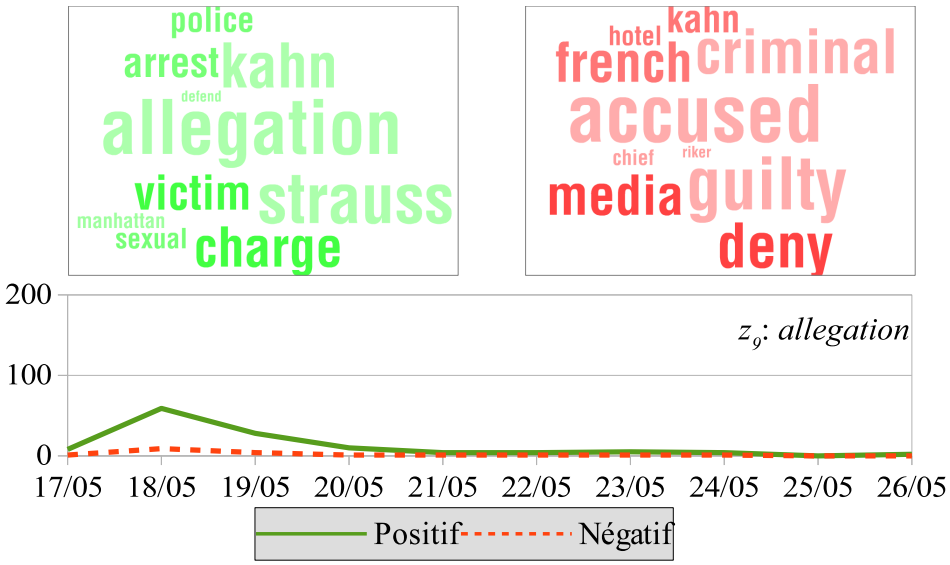


FIGURE 6.5.3 – Présentation de la thématique *allegation* (accusations).

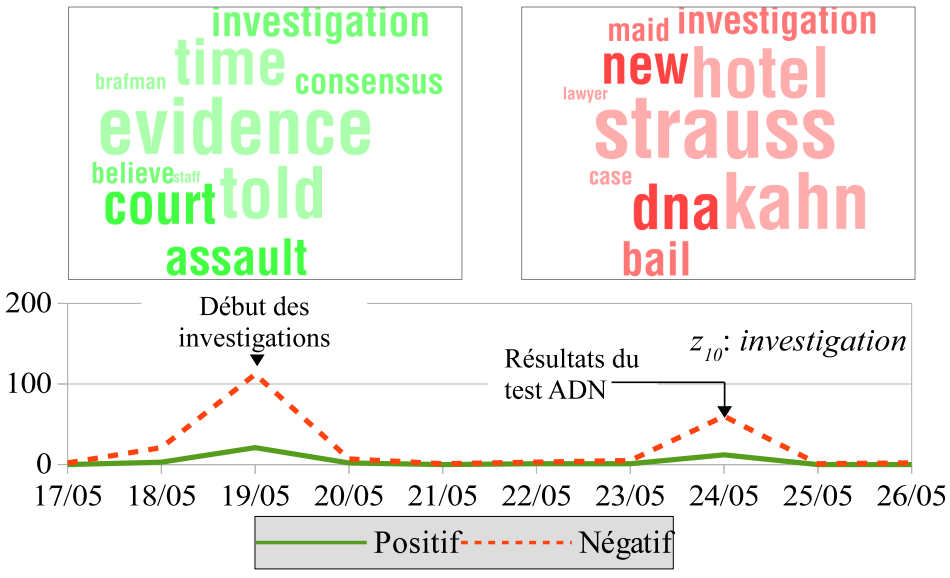


FIGURE 6.5.4 – Présentation de la thématique *investigation* (investigations).

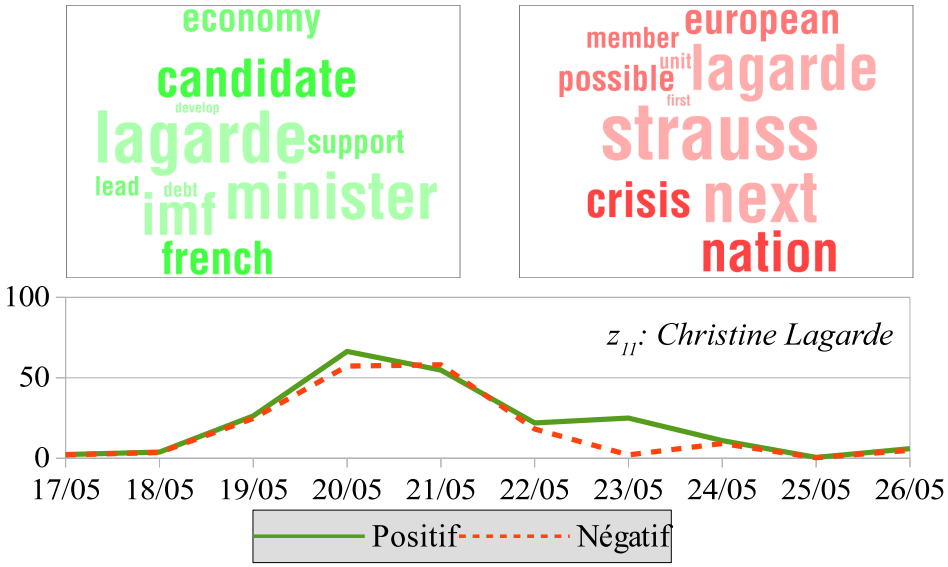


FIGURE 6.5.5 – Présentation de la thématique *Christine Lagarde* (remplaçante de DSK à la tête du FMI).

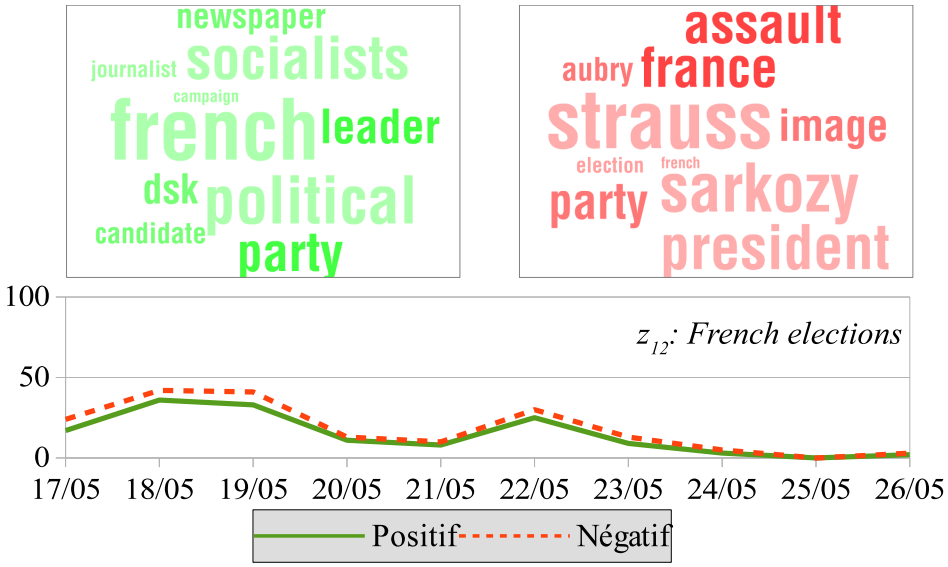


FIGURE 6.5.6 – Présentation de la thématique *French elections* (élections présidentielles françaises de 2012).

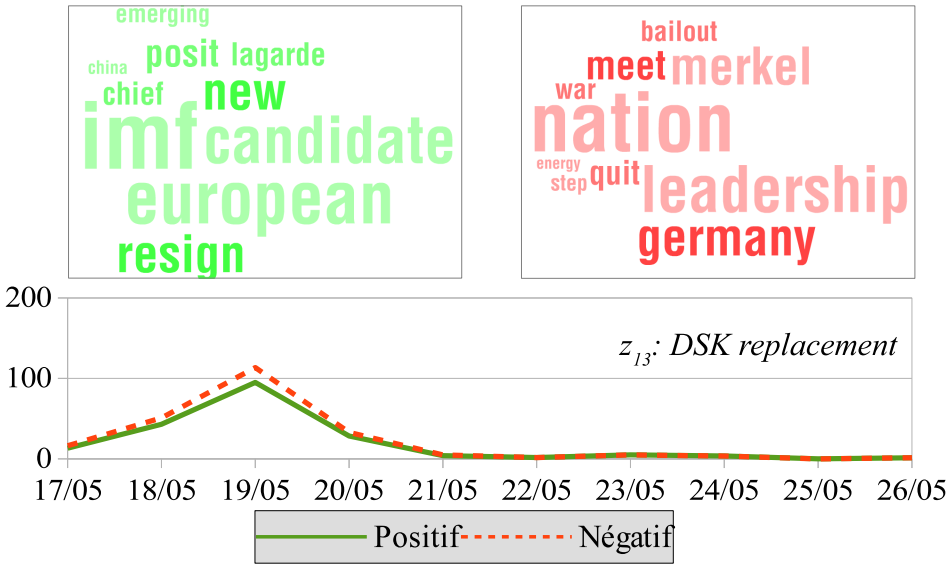


FIGURE 6.5.7 – Présentation de la thématique *replacement* (démission de DSK et son remplacement à la tête du FMI).

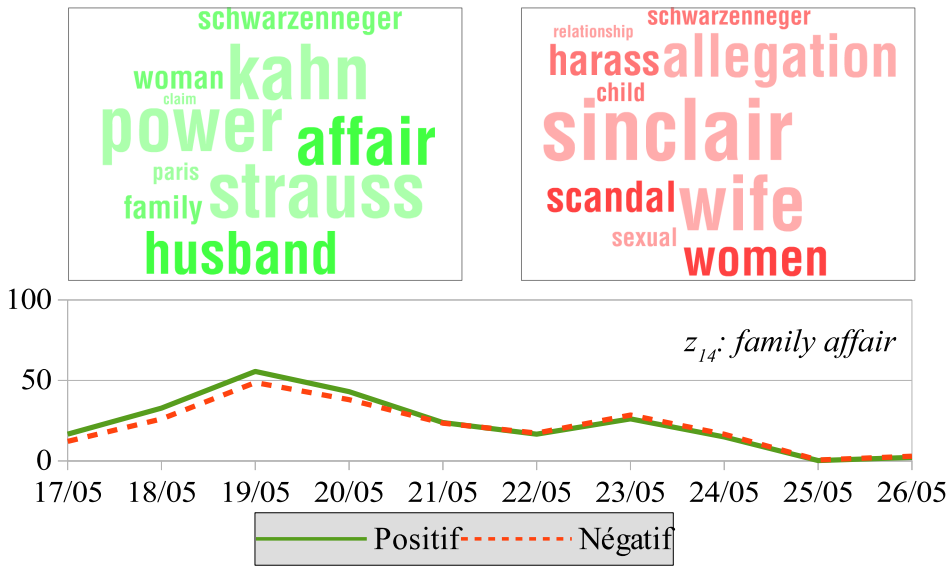


FIGURE 6.5.8 – Présentation de la thématique *family affair* (affaire de famille).

Chapitre 7

Conclusion et Perspectives

Dans cette thèse, nous nous sommes intéressés au problème de l'extraction des thématiques (de quoi parle-t-on ?) et des opinions relatives (comment en parle-t-on ?) ainsi que leur évolution temporelle. Comme nous l'avons montré, ce problème a été abordé dans la littérature à travers la proposition de plusieurs modèles probabilistes. Cependant, tous ces modèles présentent l'un ou l'autre des inconvénients suivants :

- Incapacité à caractériser les proportions d'opinions relatives aux thématiques de manière globale.
- Incapacité à caractériser l'évolution temporelle des thématiques et des opinions simultanément.

Afin d'y remédier, nous avons adopté une approche probabiliste basée sur les modèles de thématiques (*topic models*). En effet, ces méthodes permettent de prendre en compte les associations et les interactions entre les différents axes (thématiques, opinions, temps) et permettent ainsi une modélisation plus précise.

L'ensemble des travaux réalisés dans le cadre de cette thèse peut être résumé en quatre contributions principales :

1. La mesure de Vraisemblance Généralisée pour l'évaluation des méthodes d'extraction de thématiques
2. La méthode hybride pour l'analyse d'opinions
3. Le modèle TS pour l'analyse conjointe des thématiques et des opinions
4. Le modèle TTS pour l'analyse conjointe et dynamique des thématiques et des opinions.

La mesure de Vraisemblance Généralisée. La Vraisemblance Généralisée (VG) est une nouvelle mesure d'évaluation qui permet de comparer différentes méthodes d'extraction de thématiques. Ces méthodes se basent généralement sur différents paradigmes pour présenter les résultats de l'extraction, comme les distributions de probabilités, les matrices et les partitions de documents, ce qui

complique la tâche d'évaluation et de comparaison de ces résultats de manière uniforme. Afin de résoudre ce problème, nous avons proposé une méthodologie d'évaluation qui se déroule en deux étapes : nous avons d'abord défini un espace unifié et des opérations de transformation pour re-décrire les résultats dans cet espace. Ensuite, nous avons proposé la mesure VG en nous inspirant du concept de la vraisemblance largement utilisé dans la modélisation probabiliste.

Cette proposition nous a permis de comparer trois méthodes pour extraire des thématiques : LDA, NMF et FCM. Les résultats obtenus sur deux corpus différents ont donné l'avantage à LDA, suivi de NMF. Ces résultats étaient conformes à une évaluation qualitative (manuelle) de l'homogénéité et de la sémantique des thématiques extraites.

La méthode hybride pour l'analyse d'opinions. dans cette contribution, nous nous sommes intéressés au problème de sur-apprentissage rencontré avec la méthode d'apprentissage supervisé *Naive Bayes* dans un contexte de classement d'opinions selon la polarité (positive, négative et neutre). En effet, cette méthode s'est avérée particulièrement sensible à la qualité des données d'apprentissage, notamment quand ces données sont déséquilibrées ou pas suffisamment représentatives. Afin d'atténuer ce phénomène, nous avons proposé une nouvelle méthode qui combine l'apprentissage automatique et une connaissance *a priori* exprimée sous forme d'un lexique d'opinion.

Les stratégies que nous avons proposées pour combiner ces deux types de connaissances ont montré leur efficacité. En effet, sur la base des tests réalisés sur des données Twitter et des données de critiques sur le Web francophones et anglophones, notre méthode a permis d'améliorer les scores de classement.

Le modèle TS. TS (*Topic-Sentiment model*) est un modèle probabiliste que nous avons proposé dans l'objectif de caractériser les proportions d'opinions relativement à des thématiques. Contrairement aux nombreux modèles de l'état de l'art, TS permet d'extraire cette information globalement et de fournir ainsi une vue d'ensemble des associations thématiques-opinions.

Les tests effectués ont montré sa performance du modèle en terme de prédiction de l'opinion au niveau de la thématique. En effet, comparé aux modèles de la littérature, TS a donné les meilleurs taux de succès. De plus, étant basé sur la modélisation probabiliste, TS est facilement extensible afin d'intégrer d'autres dimensions d'analyse.

Le modèle TTS. TTS (*Time-aware Topic-Sentiment model*) est un modèle probabiliste que nous avons développé pour répondre à une problématique nouvelle : modéliser l'évolution temporelle des thématiques et des opinions. Le modèle TTS permet de caractériser l'évolution des thématiques et des opinions

quantitativement, c'est-à-dire en terme de volume de données. Nous avons montré à travers différents cas d'étude, notamment le cas de l'affaire DSK, comment le modèle TTS peut aider à mieux analyser ces données en faisant correspondre les résultats de l'analyse avec les événements de l'actualité. Le modèle TTS peut être utilisé pour plusieurs applications, comme l'analyse de tendances, la e-réputation et de manière plus générale l'analyse exploratoire des données textuelles selon les trois axes : thématiques, opinions et temps.

Le modèle TTS, comme d'autres modèles probabilistes, s'est montré très sensible à la valeur de l'hyperparamètre γ (paramètre responsable de l'estimation des opinions pour les thématiques). Pour remédier à ce problème, nous avons proposé une méthode automatique basée sur l'exploitation de la répartition des termes selon les polarités de l'opinion. Notre méthode ne fait usage d'aucune ressource externe sauf d'un lexique d'opinions, ce qui la rend adaptable quels que soient le domaine, la langue et la taille des données. Notre méthode pour estimer les hyperparamètres du modèle TS a montré de meilleures performances que la méthode à base de maximisation de la vraisemblance. Cette contribution est d'une grande utilité dans un contexte industriel comme le nôtre car elle permet d'automatiser le déploiement du modèle TTS et elle facilite son utilisation même sans connaissance technique de son fonctionnement.

Par ailleurs, les travaux réalisés dans cette thèse ont été intégrés dans la plateforme de veille AMIEI développée par l'entreprise AMI Software. L'intégration de ces travaux a été réalisée sous forme de nouveaux composants d'analyse avancée au sein de la plateforme AMIEI. De plus, toutes les méthodes que nous avons développées ont été testées et validées aux travers des différentes expérimentations et évaluations réalisées. La pertinence de ces méthodes a pu être validée sur des données et des cas d'usage réels issues de la plateforme de veille d'AMI Software et répondants à des problématiques concrètes posées par ses clients.

Les travaux réalisés durant cette thèse peuvent être étendus selon différents axes. D'abord, nos travaux sont basés sur l'utilisation du modèle vectoriel pour la représentation des documents. Or, cette représentation ne préserve pas les caractéristiques linguistiques du texte, comme l'ordre des mots et les relations syntaxiques. Plusieurs travaux ont montré l'intérêt des modèles linguistiques à base de n-grammes pour l'analyse d'opinions [84, 86]. Par ailleurs, certains modèles probabilistes, comme ASUM, se basent sur la phrase comme unité thématique, ce qui permet d'extraire des thématiques plus homogènes et de les associer aux opinions de manière plus précise. L'extension de nos travaux avec des connaissances linguistiques pourrait être bénéfique à la modélisation des thématiques et des opinions et contribuer à mieux analyser leurs associations.

La structure des données est un autre type de connaissance qui peut aussi enrichir la représentation des documents. En effet, certains types de données, comme les réseaux sociaux et les forums de discussion, sont structurés (par exemple sous la forme de fils de discussion ou de questions/réponses). L'exploitation de la structure des données (relations entre les documents et entre leurs créateurs) peut aider à réaliser une meilleure modélisation et fournir un résultat plus riche, par exemple, permettre la détection de communautés qui partagent les mêmes positions ou la détection des thématiques discutées par une certaine communauté [87, 97]. Pour cela, il faudrait développer des outils et des techniques pour combiner les connaissances provenant du contenu et celles provenant de la structure.

Enfin, il serait également nécessaire de poursuivre le travail de visualisation. En effet, les techniques utilisées pour visualiser les thématiques, les opinions, leurs associations et leurs évolutions sont basiques (nuage de termes, graphiques, ..). Une visualisation efficace pourrait être basée sur un processus itératif impliquant une interaction avec l'utilisateur [14, 38, 119]. Cela afin de permettre un accès à l'information d'intérêt de manière plus rapide et plus efficace.

Chapitre 8

Bibliographie

[1] Henry Anaya-Sánchez, Aurora Pons-Porrata, and Rafael Berlanga-Llavori. A new document clustering algorithm for topic discovering and labeling. In *Proceedings of the 13th Iberoamerican congress on Pattern Recognition : Progress in Pattern Recognition, Image Analysis and Applications (CIARP '08)*, pages 161–168, Havana, Cuba, 2008. Springer-Verlag.

[2] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On Finding the Natural Number of Topics with Latent Dirichlet Allocation : Some Observations. In *Proceedings of the 14th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD'10) - Volume Part I*, pages 391–402, Hyderabad, India, 2010. Springer-Verlag.

[3] Sitaram Asur and Bernardo A. Huberman. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'10)*, pages 492–499, Washington, DC, USA, 2010. IEEE Computer Society.

[4] Brigitte Bigi. Using Kullback-Leibler Distance for Text Categorization. In *Proceedings of the 25th European conference on IR research (ECIR'03)*, pages 305–319, Pisa, Italy, 2003. Springer-Verlag.

[5] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *Advances in Neural Information Processing Systems (NIPS'2003)*, Vancouver, Canada, 2003. MIT.

[6] David M. Blei and John D. Lafferty. Correlated Topic Models. In *Advances in Neural Information Processing Systems (NIPS'2006)*, volume 18, pages 147–154, Vancouver, Canada, 2006. MIT.

[7] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, pages 113–120, Pittsburgh, PA, USA, 2006. ACM.

[8] David M. Blei and Jon D. Mcauliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS'2007)*, pages 121–128, Vancouver, Canada, 2007. Curran Associates, Inc.

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research (JMLR)*, 3 :993–1022, 2003.

[10] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, pages 440–447, Prague, Czech Republic, 2007. ACL.

[11] Jordan Boyd-Graber, David M. Blei, and Xiaojin Zhu. A Topic Model for Word Sense Disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL'07)*, pages 1024–1033, Prague, Czech Republic, 2007. ACL.

[12] Soumen Chakrabarti. *Mining the Web : Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Francisco, CA, USA, 2003.

[13] Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. Reading tea leaves : How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS'2009)*, volume 31, pages 288–296, Vancouver, Canada, 2009. Curran Associates, Inc.

[14] Patricia Crossno, Andrew T. Wilson, Timothy M. Shead, Warren L. Davis IV, and Daniel M. Dunlavy. Topicview : Visual Analysis of Topic Models and their Impact on Document Clustering. *International Journal on Artificial Intelligence Tools (IJAIT)*, 22(5), 2013.

[15] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. Knowledge discovery through directed probabilistic topic models : a survey. *Frontiers of Computer Science in China (FCS)*, 4(2) :280–301, January 2010.

[16] Kushal Dave, Steve Lawrence, and David M. Pennock. 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, Budapest, Hungary, 2003. ACM.

[17] Aynur Dayanik, David D. Lewis, David Madigan, Vladimir Menkov, and Alexander Genkin. Constructing informative prior distributions from

domain knowledge in text classification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06)*, pages 493–500, Seattle, WA, USA, 2006. ACM.

[18] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science (JASIS)*, 41(6) :391–407, 1990.

[19] Mohamed Dermouche, Leila Khouas, Sabine Loudcher, Julien Velcin, and Eric Fourboul. Analyse et visualisation d ' opinions dans un cadre de veille sur le Web. In *Actes de la 15ème conference sur l'Extraction et la Gestion des Connaissances (EGC'15)*, pages 461–466, Luxembourg, Luxembourg, 2015. Hermann-Editions.

[20] Mohamed Dermouche, Leila Khouas, Julien Velcin, and Sabine Loudcher. AMI & ERIC : How to Learn with Naive Bayes and Prior Knowledge : an Application to Sentiment Analysis. In *Proceedings of : 2nd Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2 : Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 364–368, Atlanta, GA, USA, 2013. ACL.

[21] Mohamed Dermouche, Leila Khouas, Julien Velcin, and Sabine Loudcher. A Joint Model for Topic-Sentiment Modeling from Text. In *Proceedings of The 30th ACM/SIGAPP Symposium On Applied Computing (SAC'2015)*, Salamanca, Spain, 2015. ACM.

[22] Mohamed Dermouche, Julien Velcin, Leila Khouas, and Sabine Loudcher. A Joint Model for Topic-Sentiment Evolution over Time. In *Proceedings of The IEEE 14th International Conference on Data Mining (ICDM'2014)*, pages 773–778, Shenzhen, China, 2014. IEEE Computer Society.

[23] Mohamed Dermouche, Julien Velcin, Sabine Loudcher, and Leila Khouas. Une nouvelle mesure pour l'évaluation des méthodes d'extraction de thématiques : la Vraisemblance Généralisée. In *Actes de la 13ème conference sur l'Extraction et la Gestion des Connaissances (EGC'13)*, pages 317–328, Toulouse, France, 2013. Hermann-Editions.

[24] Inderjit S. Dhillon and Dharmendra S. Modha. Concept Decompositions for Large Sparse Text Data using Clustering. *Machine Learning*, 42(1-2) :143–175, 2001.

[25] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM'08)*, pages 231–240, New York, NY, USA, 2008. ACM.

[26] Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'08)*, pages 595–602, Singapore, Singapore, 2008. ACM.

[27] J.C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3) :32–57, 1973.

[28] Paul Ekman. An argument for basic emotions, 1992.

[29] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management (CIKM'05)*, pages 617–624, Bremen, Germany, 2005. ACM.

[30] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet : A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pages 417–422, Genova, IT, 2006.

[31] Yi Fang, Luo Si, Naveen Somasundaram, and Zhengtao Yu. 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, New York, NY, USA, 2012. ACM.

[32] Olivier Ferret. Approches endogène et exogène pour améliorer la segmentation thématique de documents. *Traitement Automatique des Langues (TAL), numéro spécial Discours et Document*, 47 :111–135, 2006.

[33] Eibe Frank and Remco R. Bouckaert. Naive Bayes for Text Classification with Unbalanced Classes. In *PKDD'06 Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases*, volume 4213, pages 503–510, Berlin, Germany, 2006. Springer-Verlag.

[34] Benjamin C M Fung, Ke Wang, and Martin Ester. Hierarchical Document Clustering Using Frequent Itemsets. In Daniel Barbará and Chandrika Kamath, editors, *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM'03)*, volume 30, pages 59–70, San Francisco, CA, USA, 2003. SIAM.

[35] Sara Elena Garza and Ramón Brena. Graph Local Clustering for Topic Detection in Web Collections. In *Proceedings of the 2009 Latin American Web Congress*, number June, pages 207–213, Merida, Mexico, November 2009. IEEE Computer Society.

[36] André Gohr, Alexander Hinneburg, René Schult, and Myra Spiliopoulou. Topic evolution in a stream of documents. In *Proceedings of SIAM International Conference on Data Mining (SDM'09)*, pages 859–872, Sparks, NV, USA, 2009. SIAM.

[37] Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. Learning Topics and Positions from Debatepedia. In *Proccedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*, pages 1858–1868, Seattle, WA, USA, 2013. ACL.

[38] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Tobias Hollerer, Arthur Asuncion, David Newman, and Padhraic Smyth. TopicNets : Visual Analysis of Large Text Corpora with Topic Modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(2) :23 :1–23 :26, 2012.

[39] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 :5228–5235, April 2004.

[40] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2-3) :107–145, 2001.

[41] Ali Harb, Gérard Dray, Michel Plantié, Mathieu Roche, François Trousset, and Pascal Poncelet. Web opinion mining : How to extract opinions from blogs ? In *Proceeding of International Conference on Soft Computing as Transdisciplinary Science and Technology (CSTST'08)*, pages 211–217, New York, NY, USA, 2008. ACM.

[42] Donna Harman. Overview of the first TREC conference. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'93)*, pages 36–47. ACM, 1993.

[43] Vasileios Hatzivassiloglou and Kathleen R Mckeown. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 8th conference of the European chapter of the Association for Computational Linguistics (EACL'97)*, pages 174–181, Madrid, Spain, 1997. ACL.

[44] Thomas Hofmann. 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, Berkeley, CA, USA, 1999. ACM.

[45] Minqing Hu and Bing Liu. 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, Seattle, WA, USA, 2004. ACM.

[46] Jonathan Huang. Maximum likelihood estimation of Dirichlet distribution parameters. Technical report, 2005.

[47] Ioana Hulpus, Conor Hayes, Marcel Karnstedt, and Derek Greene. Unsupervised graph-based topic labelling using DBpedia. In *Proceedings of the 6th ACM international conference on Web search and data mining (WSDM'13)*, pages 465–474, New York, NY, USA, 2013. ACM.

[48] Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (HLT'11) - Volume 1*, pages 151–160, Portland, Oregon, 2011. ACL.

[49] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM'11)*, pages 815–824, Hong Kong, China, 2011. ACM.

[50] Jaap Kamps, Maarten Marx, Robert J. Mokken, and Maarten de Rijke. Using wordnet to measure semantic orientations of adjectives. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC-04)*, pages 1115–1118, Lisbon, PT, 2004.

[51] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contexual valence shifters. *Computational Intelligence*, 22(2) :110–125, May 2006.

[52] Suin Kim, Jianwen Zhang, Zheng Chen, Alice Oh, and Shixia Liu. A Hierarchical Aspect-Sentiment Model for Online Reviews. In *Proceedings of The 27th AAAI Conference on Artificial Intelligence (AAAI'13)*, pages 526–533, Bellevue, WA, USA, 2013. AAAI Press.

[53] Jon M. Kleinberg, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a graph : measurements , models , and methods. In *Proceedings of the 5th Annual International Conference on Computing and Combinatorics (COCOON'99)*, pages 1–17, Tokyo, Japan, 1999. Springer.

[54] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis : The good the bad and the OMG ! In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM'11)*, pages 538–541, Barcelona, Spain, 2011. The AAAI Press.

[55] Simon Lacoste-julien, Fei Sha, and Michael I. Jordan. DiscLDA : Discriminative Learning for Dimensionality Reduction and Classification. In *Advances in Neural Information Processing Systems (NIPS'2008)*, pages 897–904, Vancouver, Canada, 2008. Curran Associates, Inc.

[56] Victor Lavrenko and W. Bruce Croft. Relevance-Based Language Models. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in Information Retrieval (SIGIR'01)*, pages 120–127, New Orleans, LA, USA, 2001. ACM.

[57] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–91, October 1999.

[58] Daniel D. Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems (NIPS'2000)*, volume 13, pages 556–562, Denver, CO, USA, 2000. MIT Press.

[59] Chengtao Li, Jianwen Zhang, Jian-tao Sun, and Zheng Chen. Sentiment Topic Model with Decomposed Prior. In *Proceedings of 2013 SIAM International Conference on Data Mining (SDM'13)*, pages 767–776, Austin, TX, USA, 2013. SIAM.

[60] Fangtao Li, Minlie Huang, and Xiaoyan Zhu. Sentiment analysis with global topics and local dependency. *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI'10)*, pages 1371–1376, 2010.

[61] Linlin Li, Benjamin Roth, and Caroline Sporleder. Topic Models for Word Sense Disambiguation and Token-Based Idiom Detection. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1138–1147, Uppsala, Sweden, 2010. ACL.

[62] Wei Li and Andrew Mccallum. Pachinko Allocation : DAG-Structured Mixture Models of Topic Correlations. In *Proceedings of the 23rd international conference on Machine learning (ICML'06)*, pages 577–584, Pittsburgh, PA, USA, 2006. ACM.

[63] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM'09)*, pages 375–384, Hong Kong, China, 2009. ACM.

[64] Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 24(6) :1134–1145, June 2012.

[65] Kang Liu, Liheng Xu, and Jun Zhao. Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. In *314 Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL'14)*, pages 314–324, Baltimore, MD, USA, 2014. ACL.

[66] Yabing Liu, C Kliman-Silver, and Alan Mislove. The Tweets They are a-Changin' : Evolution of Twitter Users and Behavior. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14)*, pages 305–314, Ann Arbor, MI, USA, 2014. AAAI Press.

[67] Davide Magatti, Silvia Calegari, Davide Ciucci, and Fabio Stella. Automatic labeling of topics. In *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications (ISDA'09)*, pages 1227–1232, Pisa, Italy, 2009. IEEE Computer Society.

[68] Sigrid Maurel, Paolo Curtoni, and Luca Dini. Classification d'opinions par méthodes symbolique, statistique et hybride. In *Actes du 3ème DEfi Fouille de Textes (DEFT'07)*, pages 121–127, Grenoble, France, 2007.

[69] Andrew K. McCallum. MALLET : A Machine Learning for Language Toolkit, 2002.

[70] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 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, Banff, Canada, 2007. ACM.

[71] Qiaozhu Mei, Xuehua Shen, and Chengxiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'07)*, pages 490–499, San Jose, CA, USA, 2007. ACM.

[72] Prem Melville, Wojciech Gryc, and Richard D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*, pages 1275–1284, New York, NY, USA, 2009. ACM Press.

[73] Cong Meng, Mian Zhang, and Wenqiong Guo. Evolution of Movie Topics Over Time. Technical report, 2012.

[74] George A. Miller. WordNet : a lexical database for English. *Communications of the ACM*, 38(11) :39–41, 1995.

[75] Thomas Minka and John Lafferty. Expectation-Propagation for the Generative Aspect Model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence (UAI'02)*, pages 352–359, Alberta, Canada, 2002. Morgan Kaufmann Publishers Inc.

[76] Thomas P. Minka. Estimating a Dirichlet distribution. Technical Report 8, MIT, 2003.

[77] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi, and Toshikazu Fukushima. Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)*, pages 341–349, Edmonton, Canada, 2002. ACM.

[78] Claudiu Cristian Musat, Julien Velcin, Stefan Trausan-matu, and Marian-andrei Rizoiu. Improving Topic Evaluation Using Conceptual Knowledge. In *Proceedings of the Twenty-Second international joint*

conference on Artificial Intelligence (IJCAI'11), pages 1866–1871, Barcelona, Spain, 2011. AAAI Press.

[79] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, and Anna Becker. *Réseaux Bayésiens*. Eyrolles, 2007.

[80] David J. Newman and Sharon Block. Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper. *Journal of the American Society for Information Science and Technology (JASIST)*, 57(6) :753–767, 2006.

[81] Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. Examining the Role of Linguistic Knowledge Sources in the Automatic Identification and Classification of Reviews. In *Proceedings of the COLING/ACL 2006 Main conference poster sessions*, pages 611–618, Sydney, Australia, 2006. ACL.

[82] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 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*, Washington, DC, USA, 2010.

[83] Alexander Pak and Patrick Paroubek. Construction d'un lexique affectif pour le français à partir de Twitter. In *Actes de la 17ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, pages 19–23, Montréal, Canada, 2010.

[84] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pages 1320–1326, Valletta, Malta, 2010. ELRA.

[85] Bo Pang and Lillian Lee. 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, Barcelona, Catalonia, Spain, 2004. ACL.

[86] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP'02)*, pages 79–86, Philadelphia, PA, USA, 2002. ACL.

[87] Nishith Pathak, Colin DeLong, Arindam Banerjee, and Kendrick Erickson. Social topic models for community extraction. In *The 2nd SNA-KDD Workshop (SNA-KDD'08)*, Las Vegas, Nevada, USA, 2008.

[88] Aurora Pons-Porrata, Rafael Berlanga-Llavori, and José Ruiz-Shulcloper. Building a hierarchy of events and topics for newspaper digital libraries. In *Proceedings of the 25th European Conference on Information Retrieval research (ECIR'03)*, pages 588–596, Pisa, Italy, 2003. Springer-Verlag.

[89] Ana-Maria Popescu and Oren Etzioni. Extracting Product Features and Opinions from Reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*, pages 339–346, Vancouver, Canada, 2005. ACL.

[90] Martin F. Porter. An algorithm for suffix stripping. In *Readings in information retrieval*, number 3, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997.

[91] Dorian Pyle. *Data Preparation for Data Mining.* Morgan Kaufmann Publishers Inc., 1999.

[92] Core Team R. R : A Language and Environment for Statistical Computing, 2012.

[93] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, pages 616–623, Washington, DC, USA, 2003. AAAI Press.

[94] Antonio Reyes and Paolo Rosso. Mining subjective knowledge from customer reviews : a specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA'11)*, pages 118–124, Portland, OR, USA, 2011. ACL.

[95] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processin (EMNLP'13)*, pages 704–714, Seattle, WA, USA, 2013. ACL.

[96] Stephen E. Robertson and Karen Sparck Jones. Relevance Weighting of Search Term. In *Document Retrieval Systems*, pages 143–160. Taylor Graham Publishing, 1988.

[97] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI'04)*, pages 487–494, Banff, Canada, 2004. AUAI Press.

[98] Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. Finding the Sources and Targets of Subjective Expressions. In *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*, number 2, Marrakech, Morocco, 2008. ELRA.

[99] Gerard Salton, Andrew K. C. Wong, and Chung S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11) :613–620, 1975.

[100] R.E. Schapire, M. Rochery, M. Rahim, and N. Gupta. Incorporating prior knowledge into boosting. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML'02)*, pages 538–545, Sydney, Australia, 2002. Morgan Kaufmann Publishers Inc.

[101] Jude W. Shavlik. A Framework for Combining Symbolic and Neural Learning. *Machine Learning*, 14(3) :321–331, 1994.

[102] Philipp Singer, Claudia Wagner, and Markus Strohmaier. Factors influencing the co-evolution of social and content networks in online social media. In *Proceedings of the 2011 international conference on Modeling and Mining Ubiquitous Social Media (MSM'11)*, pages 40–59, Boston, MA, USA, 2011. Springer-Verlag.

[103] AMI Software. AMI Enterprise Intelligence : Description de Produit Logiciel. Technical Report 7.0, AMI Software, 2015.

[104] Gamgarn Somprasertsri and Pattarachai Lalitrojwong. Mining Feature-Opinion in Online Customer Reviews for Opinion Summarization. *Journal of Universal Computer Science*, 16(6) :938–955, 2010.

[105] Michael Steinbach, George Karypis, and Vipin Kumar. A comparison of document clustering techniques. In *Proceedings of KDD-2000 workshop on text mining*, pages 1–20, Boston, MA, USA, 2000. ACM.

[106] Veselin Stoyanov and Claire Cardie. Annotating topics of opinions. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)*, number 1, pages 2–6, Marrakech, Morocco, 2008. ELRA.

[107] Veselin Stoyanov and Claire Cardie. Topic Identification for Fine-Grained Opinion Analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING'08) - Volume 1*, pages 817–824, Manchester, United Kingdom, 2008. ACL.

[108] Amine Trabelsi and Osmar R. Zaiane. A Joint Topic Viewpoint Model for Contention Analysis. In *Proceedings of the 19th International Conference on Applications of Natural Language to Information Systems (NLDB'14)*, pages 114–125, Montpellier, France, 2014. Springer.

[109] Amine Trabelsi and Osmar R. Zaiane. Mining Contentious Documents Using an Unsupervised Topic Model Based Approach. In *IEEE International Conference on Data Mining (ICDM'14)*, pages 550–559, Shenzhen, China, 2014. IEEE.

[110] Oren Tsur, Dmitry Davidov, and Ari Rappoport. ICWSM - A Great Catchy Name : Semi-Supervised Recognition of Sarcastic Sentences in Product Reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM'10)*, pages 162–169, Washington, DC, USA, 2010. AAAI Press.

[111] Peter D. Turney and Michael L. Littman. Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4) :315–346, 2003.

[112] Marc Vincent and Grégoire Winterstein. Construction et exploitation d'un corpus français pour l'analyse de sentiment. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'13)*, number 2007, pages 764–771, Les Sables d'Olonne, France, 2013. ATALA.

[113] Hanna M. Wallach, David Mimno, and Andrew Mccallum. Rethinking LDA : Why Priors Matter. In *Advances in Neural Information Processing Systems (NIPS'2009)*, pages 1973–1981, Vancouver, Canada, 2009. Curran Associates, Inc.

[114] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 1105–1112, Montreal, Canada, 2009. ACM.

[115] Hao Wang and Martin Ester. A Sentiment-aligned Topic Model for Product Aspect Rating Prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*, pages 1192–1202, Doha, Qatar, 2014. ACL.

[116] Xuerui Wang and Andrew McCallum. Topics over time : a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'06)*, pages 424–433, Philadelphia, PA, USA, 2006. ACM.

[117] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and text. In *Proceedings of the 3rd international workshop on Link discovery (LinkKDD'05)*, pages 28–35, Chicago, IL, USA, 2005. ACM.

[118] Christian Wartena and Rogier Brussee. Topic Detection by Clustering Keywords. In *Proceedings of the 2008 19th International Conference on Database and Expert Systems Application (DEXA'08)*, pages 54–58, Turin, Italy, 2008. IEEE Computer Society.

[119] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. TIARA : A Visual Exploratory Text Analytic System. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'10)*, pages 153–162, Washington, DC, USA, 2010. ACM.

[120] Dawid Weiss. *Descriptive clustering as a method for exploring text collections*. PhD thesis, Poznan University of Technology, Poland, 2006.

[121] Sinead Williamson, Trumpington Street, Chong Wang, Katherine A. Heller, and David M. Blei. The IBP Compound Dirichlet Process and its Application to Focused Topic Modeling. In *International Conference on Machine Learning (ICML'10)*, pages 1151–1158, Haifa, Israel, 2010. Omnipress.

[122] Theresa Wilson, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter. SemEval-2013 Task 2 : Sentiment Analysis in Twitter. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval'13)*, pages 312–320, Atlanta, GA, USA, 2013. ACL.

[123] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, pages 347–354, Vancouver, Canada, 2005. ACL.

[124] Xiaoyun Wu and Rohini Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04)*, pages 326–333, Seattle, WA, USA, 2004. ACM.

[125] Liheng Xu, Kang Liu, Siwei Lai, Yubo Chen, and Jun Zhao. Mining Opinion Words and Opinion Targets in a Two-Stage Framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL'13)*, pages 1764–1773, Sofia, Bulgaria, 2013. ACL.

[126] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR'03)*, pages 267–273, Toronto, Canada, 2003. ACM.

[127] Xing Yi and James Allan. A Comparative Study of Utilizing Topic Models for Information Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR'09)*, pages 29–41, Toulouse, France, 2009. Springer-Verlag.

[128] ChengXiang Zhai and John D. Lafferty. Model-based Feedback In The Language Modeling Approach To Information Retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01)*, pages 403–410, Atlanta, GA, USA, 2001. ACM.

[129] Chengxiang Zhai and John D. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2) :179–214, 2004.

[130] Chengzhi Zhang, Huilin Wang, Yao Liu, and Hongjiao Xu. Document Clustering Description Extraction and Its Application. In *Proceedings*

of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy (ICCPOL'09), pages 370–377, Hong Kong, China, 2009. Springer-Verlag.

[131] Minjie Zheng, Chaorong Wu, Yue Liu, Xiangwen Liao, and Guolong Chen. Topic Sentiment Trend Model : Modeling facets and sentiment dynamics. In *Proceedings of the 2nd IEEE International Conference on Computer Science and Automation Engineering (CSAE'12)*, number 1, pages 651–657, Zhangjiajie, China, 2012. IEEE.

Table des matières

Table des figures

Liste des tableaux

Annexe A

Liste des publications

• Mohamed Dermouche, Leila Khouas, Julien Velcin, and Sabine Loudcher. A Joint Model for Topic-Sentiment Modeling from Text. In Proceedings of The 30th ACM/SIGAPP Symposium On Applied Computing (**SAC'2015**), Salamanca, Spain, 2015. ACM.

• Mohamed Dermouche, Leila Khouas, Sabine Loudcher, Julien Velcin, and Eric Fourboul. Analyse et visualisation d'opinions dans un cadre de veille sur le Web. In Actes de La 15ème Conférence Sur l'Extraction et La Gestion Des Connaissances (**EGC'15**), pages 461–466, Luxembourg, Luxembourg, 2015. Hermann-Editions.

• Mohamed Dermouche, Julien Velcin, Leila Khouas, and Sabine Loudcher. A Joint Model for Topic-Sentiment Evolution over Time. In Proceedings of The IEEE 14th International Conference on Data Mining (**ICDM'2014**), pages 773–778, Shenzhen, China, 2014. IEEE Computer Society.

• Mohamed Dermouche, Leila Khouas, Julien Velcin, and Sabine Loudcher. AMI & ERIC : How to Learn with Naive Bayes and Prior Knowledge : An Application to Sentiment Analysis. In Proceedings of the 7th International Workshop on Semantic Evaluation (**SemEval'2013**), volume 2, pages 364–368, Atlanta, GA, USA, 2013. ACL.

• Mohamed Dermouche, Julien Velcin, Sabine Loudcher, and Leila Khouas. Une nouvelle mesure pour l'évaluation des méthodes d'extraction de thématiques : la Vraisemblance Généralisée. In Actes de La 13ème Conférence Sur l'Extraction et La Gestion Des Connaissances (**EGC'13**), pages 317–328, Toulouse, France, 2013. Hermann-Editions.

Annexe B

Glossaire

Classification (*clustering*) : tâche de fouille de données qui consiste à faire émerger des groupes de données par apprentissage non supervisé.

Classement (*classification, categorization*) : tâche de fouille de données qui consiste à affecter des données à des classes sur la base d'un modèle construit par apprentissage supervisé.

Centroïde : dans une classe de données, le centroïde est le vecteur moyen de toutes les données de la classe.

Corpus : ensemble de documents textuels sous format numérique.

Lexique d'opinion : dans ce manuscrit, un lexique d'opinion désigne une liste de termes polarisés où chaque terme est caractérisé par la polarité qu'il exprime.

Mots vides (*stopwords*) : mots outils spécifique à une langue, par exemple `le, et, cela`, etc.

Polarité d'opinion : modalité d'opinion, généralement positive, négative ou neutre.

Terme : attribut d'un document (mot simple, n-gramme, concept, etc.).

Thématique (*topic*) : une thématique est un sujet unique et clairement identifiable dans un ou plusieurs documents.