Influence des marqueurs multi-polaires dépendant du domaine pour la fouille d'opinion au niveau du texte

Morgane Marchand[1, 2] Romaric Besançon[1] Olivier Mesnard[1] Anne Vilnat[2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus
Centre Nano-Innov Saclay, 91191 Gif-sur-Yvette Cedex
(2) LIMSI-CNRS, Université Paris-Sud, 91403 Orsay Cedex
morgane.marchand@cea.fr, romaric.besancon@cea.fr, olivier.mesnard@cea.fr, anne.vilnat@limsi.fr

Résumé. Les méthodes de détection automatique de l'opinion dans des textes s'appuient sur l'association d'une polarité d'opinion aux mots des textes, par lexique ou par apprentissage. Or, certains mots ont des polarités qui peuvent varier selon le domaine thématique du texte. Nous proposons dans cet article une étude des mots ou groupes de mots marqueurs d'opinion au niveau du texte et qui ont une polarité changeante en fonction du domaine. Les expériences, effectuées à la fois sur des corpus français et anglais, montrent que la prise en compte de ces marqueurs permet d'améliorer de manière significative la classification de l'opinion au niveau du texte lors de l'adaptation d'un domaine source à un domaine cible. Nous montrons également que ces marqueurs peuvent être utiles, de manière limitée, lorsque l'on est en présence d'un mélange de domaines. Si les domaines ne sont pas explicites, utiliser une séparation automatique des documents permet d'obtenir les mêmes améliorations.

Abstract. In this article, we propose a study on the words or multi-words which are good indicators of the opinion polarity of a text but have different polarity depending on the domain. We have performed experiments on French and English corpora, which show that taking these multi-polarity words into account improve the opinion classification at text level in a domain adaptation framework. We also show that these words are useful when the corpus contains several domains. If these domains are not explicit, using a automatic domain characterization (e.g. wich Topic Modeling approaches) allows to achieve the same results.

Mots-clés : Fouille d'opinion, adaptation au domaine, marqueurs multi-polaires.

Keywords: Opinion mining, domain adaptation, multi-polarity markers.

Introduction

Avec l'avènement du web 2.0, la manière dont les personnes expriment leur opinion a beaucoup changée : nous postons des critiques de produits de consommation sur des sites marchands et exposons nos points de vue sur presque tous les sujets, sur des forums, des groupes de discussion ou des blogs. Tout cela constitue une importante source d'information avec de nombreuses applications. C'est pourquoi, au cours des dernières années, de nombreux travaux ont pris pour objet la fouille d'opinion. Si beaucoup de ces travaux se focalisent sur la caractérisation de l'opinion sur un corpus donné, qui est souvent spécifique à un domaine, l'étude des mots qui n'indiquent pas la même opinion d'un domaine à l'autre est moins fréquente. Certains mots peuvent en effet changer de polarité entre deux domaines (Navigli, 2012; Yoshida *et al.*, 2011). Par exemple, le mot "retourner" a une connotation positive dans la phrase "Je n'en peux plus d'attendre pour retourner à mon livre !". Mais il exprime généralement une opinion très négative s'il est employé pour parler d'un appareil électronique, comme dans "J'ai dû retourner au magasin". Ce phénomène peut apparaître même lorsque les domaines sont très proches : "J'étais mort de rire" est bon signe pour un film comique mais pas pour un film d'horreur. Dans cet article, les mots ou groupes de mots sujets à ce phénomène sont appelés des "marqueurs multi-polaires". Si on ne repère pas de tels mots lors d'une tâche de classification automatique de l'opinion, ils peuvent conduire à des erreurs de classification (Wilson *et al.*, 2009). Nous proposons ici une étude de ces marqueurs, en montrant l'apport potentiel de leur prise en compte pour l'adaptation d'un domaine à un autre ainsi que pour la détection d'opinion en domaine ouvert.
Dans une première partie, nous explicitons le concept de marqueur multi-polaire et le comparons avec les autres concepts présents dans l'état de l'art. Nous présentons ensuite la méthode utilisée pour détecter ces marqueurs, ainsi qu'une clas-

sification de la nature de ces marqueurs. Dans les parties suivantes, nous étudions l'influence des ces marqueurs sur la performance des classifieurs automatiques d'opinion lors du transfert d'un domaine source à un domaine cible, ainsi que sur la détection d'opinion sur dans corpus multi-domaines et des corpus en domaine ouvert.

1 Concept et état de l'art

Subjectivité, polarité et domaines

Les expressions subjectives sont des mots ou des groupes de mots utilisés pour exprimer des états mentaux comme la spéculation, l'évaluation, le sentiment ou la conviction (Wiebe *et al.*, 2005; Wiebe & Mihalcea, 2006; Wilson, 2008; Akkaya *et al.*, 2009a). Ils sont appelés "état privés", c'est à dire que ce sont des états internes qui ne peuvent pas être directement observés par les autres (Quirk & Crystal, 1985). La polarité d'un mot ou d'un sens particulier d'un mot, au contraire, fait référence à l'opinion positive ou négative qu'a un agent sur un objet particulier. Ces deux notions ne sont bien sûr pas indépendantes et la plupart des sens subjectifs des mots ont une polarité claire. Néanmoins, une expression polarisée peut également apparaître dans un contexte neutre (Wilson *et al.*, 2009). De plus, une polarité peut être associée à des mots ou des sens de mots objectifs. (Su & Markert, 2008) donnent l'exemple du mot *tuberculose* : ce mot ne décrit pas un état privé, on peut le vérifier de manière objective et sa présence dans une phrase ne force pas cette dernière à être porteuse d'opinion. Mais pour la plupart des gens, ce mot porte tout de même une forte connotation négative. Comme (Su & Markert, 2008), nous ne considérons pas que le fait d'être polaire soit réservé aux mots ou expressions ayant été au préalable classés comme subjectifs.
La polarité d'un mot ou d'une expression peut de plus varier en fonction du contexte. Depuis quelques années, l'intérêt pour lever l'ambiguïté sur la polarité des mots ambigus s'est amplifié (Wu & Jin, 2010). Presque tous les schémas d'annotation existant pour la polarité permettent de noter cette ambiguïté (Su & Markert, 2008; Wilson *et al.*, 2005). Nous nous intéressons ici spécifiquement aux variations de polarité dues au domaine du texte, c'est à dire à son type de sujet. Dans leur travail sur la polarité contextuelle, (Wilson *et al.*, 2005) incluent le sujet et le domaine comme causes possibles de variation de polarité. De plus, (Su & Markert, 2008) remarquent dans leur étude que des préférences de polarité existent selon le domaine ou le sujet du texte. Leur corpus contient 32,5 % de mots à la polarité ambiguë et la simple désambiguïsation de sens ne parvient pas à résoudre complètement cette ambiguïté. Dans (Takamura *et al.*, 2006, 2007), les auteurs proposent une méthode utilisant un modèle avec variable latente et réseau lexical pour déterminer l'orientation de paires adjectif+nom. Ils remarquent que si l'adjectif est ambigu, la classification est plus difficile.

Les marqueurs multi-polaires au niveau du texte

Dans cette étude, nous nous intéressons aux mots ou expressions (subjectifs tout comme objectifs) qui, de manière récurrente dans un domaine particulier, sont des indicateurs de l'opinion de l'auteur sur l'objet du texte. Tout comme pour l'exemple de la tuberculose, beaucoup de mots auxquels nous nous intéressons ne vont pas avoir de polarité intrinsèque mais peuvent apparaître dans des contextes récurrents de connotation polaire pour un domaine particulier.
Ce travail est proche des concepts de polarité contextuelle ou ciblée (Wilson *et al.*, 2005, 2009; Fahrni & Klenner, 2008). (Fahrni & Klenner, 2008) se focalisent sur la détermination de la polarité ciblée des adjectifs. Un nom spécifique à un domaine est souvent modifié par un adjectif qualificatif. D'après les auteurs, les adjectifs n'ont pas de polarité *a priori* mais une polarité ciblée. Dans certains cas, un même adjectif peut changer de polarité en fonction du nom qu'il accompagne. Les auteurs utilisent Wikipédia pour la détection automatique des mots qui peuvent potentiellement être la cible d'une opinion pour un domaine donné. Une méthode de *bootstrap* est ensuite utilisée afin de déterminer la polarité ciblée des adjectifs associés à ces mots. Ils obtiennent de bons résultats mais s'intéressent uniquement aux adjectifs. (Wilson *et al.*, 2005), quant à eux, ne se restreignent pas aux adjectifs mais travaillent uniquement sur des segments de texte contenant des mots prédéterminés (des mots d'un lexique ayant au moins un sens subjectif). Ils se placent au niveau du segment et déterminent d'abord si une expression est neutre ou polaire avant de désambiguïser la polarité des expressions polaires en utilisant des règles manuelles et des traits structurels. Leur lexique couvre 75 % des segments polaires de leur corpus.
Pour notre étude, nous ne présumons pas des mots ou des expressions qui sont porteurs ou non d'information polaire. Nous avons donc choisi de les sélectionner automatiquement et de les classifier en une seule étape.
De plus, nous nous intéressons dans cet article à l'influence des mots ou expressions à polarité ambiguë (que nous appellerons marqueurs multi-polaires) sur la valeur de la polarité du texte entier. Beaucoup de travaux utilisent un lexique donnant la polarité *a priori* des mots. Ces lexiques sont souvent construits en étendant un petit lexique initial, soit en tirant

parti des conjonctions comme et/mais (Hatzivassiloglou & McKeown, 1997), soit en mesurant la co-occurrence entre mots dans un corpus ou à l'aide de moteurs de recherche (Turney & Littman, 2002). D'autres travaux améliorent un lexique déjà pré-existant, par exemple en pondérant les différentes polarités possibles d'un mot en fonction du domaine (Choi & Cardie, 2009). Ces lexiques particuliers peuvent alors être utilisés dans des classifieurs à base de règles pour classer la polarité des textes entiers (Ding *et al.*, 2008). Les études au niveau du texte utilisant des classifieurs à base de corpus s'intéressent, quant à elles, principalement à la représentation des données (Glorot *et al.*, 2011; Huang & Yates, 2012). L'erreur d'adaptation d'un classifieur dépend en effet de sa performance sur le domaine source ainsi que de la distance entre les distributions des mots dans les domaines source et cible (Ben-David *et al.*, 2007). Avec une bonne projection, un lien peut être établi entre les mots du domaine cible qui n'existent pas dans le domaine source et les autres mots (Pan *et al.*, 2010; Blitzer *et al.*, 2007). Cependant, si un mot a une polarité différente dans le domaine source et le domaine cible, cela va introduire une erreur d'adaptation. Ainsi, la détection des marqueurs multi-polaires est complémentaire à ces approches et leurs améliorations respectives peuvent être combinées.

2 Caractérisation des marqueurs multi-polaires

2.1 Méthode de détection des marqueurs multi-polaires

Nous nous intéressons donc aux marqueurs de polarité d'opinion au niveau du texte, dont la polarité est changeante avec le domaine. Le repérage automatique de ces mots ou expressions se fait par apprentissage, en utilisant des corpus de textes issus de différents domaines et annotés globalement en fonction de leur polarité sur un axe positif-négatif. Plus précisément, pour chaque mot apparaissant dans plusieurs corpus, nous regardons si sa distribution dans les critiques positives et négatives est statistiquement différente selon les domaines [1]. La caractérisation de cette différence statistique est établie par un test du χ^2 avec un risque de première espèce (i.e. risque de faux positif) de 1 %.

2.2 Utilisation des marqueurs multi-polaires lors d'un transfert

Une fois les marqueurs multi-polaires détectés, ils peuvent servir à améliorer la classification d'opinion lors d'un transfert d'un domaine source à un domaine cible. Lors de cette tâche, un classifieur d'opinion est automatiquement appris sur le corpus source annoté avant d'être utilisé sur un domaine cible. Afin de prendre en compte l'information apportée par les marqueurs d'opinion multi-polaires détectés, nous proposons de modifier les corpus source et cible avant l'entraînement du classifieur. Nous proposons deux types de modifications différentes :

En distinguant les mots Chaque marqueur multi-polaire est différencié selon le domaine : il est remplacé par le trait *marqueur_Source* dans le corpus d'entraînement (du domaine source) et par *marqueur_Cible* dans le corpus de test (du domaine cible). Ainsi, l'erreur de transfert sur les marqueurs sélectionnés est évitée.

En enlevant les mots Chaque marqueur multi-polaire est tout simplement retiré, à la fois du corpus d'entraînement et du corpus de test.

2.3 Classification des marqueurs multi-polaires

Les changements de polarité que l'on observe dans les textes, peuvent être liés à des phénomènes linguistiques ou contextuels différents. Nous en proposons la classification suivante :

Changement de sens La multi-polarité d'un mot peut être liée à sa polysémie. Dans "*I had to return my phone to the store*"ou "*I can't wait to return to my book*", le mot *return* a une polarité différente car il s'agit de deux sens différents. Dans ce cas, utiliser une méthode de désambiguïsation de sens ou de subjectivité comme dans (Akkaya *et al.*, 2009b) peut être utile.

Qualité relative Certains adjectifs ou qualificatifs sans polarité *a priori* peuvent être positifs ou négatifs en fonction de l'objet qu'ils qualifient (Fahrni & Klenner, 2008). Être "imprévisible" est un qualificatif positif pour un scénario de film mais négatif pour un logiciel.

1. Certains mots peuvent changer de polarité à l'intérieur du même domaine mais nous ne nous intéressons ici qu'à la polarité au niveau global.

Orientation morale et politique de l'auteur Certains mots peuvent changer de polarité en fonction de l'opinion de l'auteur. Cela concerne souvent les termes politiques (par exemple "capitalisme").

Comparaison Les opinions comparatives ("meilleur que...") sont difficiles à prendre en compte car il faut alors savoir quelle partie de la comparaison est l'objet principal du texte. Des travaux sont consacrés à ce problème spécifique (Ganapathibhotla & Liu, 2008). Nous avons pu détecter des habitudes générales dans nos différents corpus. Pour certains, l'objet de la critique est dans une très grande majorité à la première place de la comparaison. Dans d'autres, c'est le contraire. Il peut cependant être délicat de vérifier qu'il s'agit bien d'un phénomène global propre à un domaine et non un biais de corpus.

Aspect temporel La polarité de certains mots peut être connectée à une information temporelle. Par exemple, "*I loved this book*" est positif mais "*I loved this camera*" est habituellement négatif car l'objet ne fonctionne en général plus. Ainsi, "*I loved*" est le signe d'une opinion négative lorsque l'on parle d'objets électroniques mais la forme au présent, "*I love*", reste positive.

Biais de corpus Un changement de polarité peut être dû à un biais de corpus. Par exemple, si beaucoup de monde est d'accord pour dire que le film *Superman* est une adaptation peu réussie de la bande dessinée classique, le mot *Superman* risque d'être associé à une polarité fortement négative dans un corpus dédié aux critiques de films.

Pour certaines de ces catégories, des traitements spécifiques existent, comme la désambiguïsation de sens ou les travaux sur les opinions comparatives. Pour d'autres, il n'existe pas de traitement usuel. C'est pourquoi étudier ces mots multi-polaires est une nécessité.
Une annotation manuelle est actuellement en cours afin d'étudier la répartition des marqueurs multi-polaires dans ces différentes classes. Nous nous attacherons notamment à comparer les phénomènes observés sur l'anglais et le français afin d'expliquer plus en détail les différences observées sur les résultats.

3 Impact des marqueurs multi-polaires pour l'adaptation au domaine

3.1 Extraction des marqueurs multi-polaires

Nous avons réalisé la détection de marqueurs multipolaires pour l'anglais et le français. Pour l'anglais, nous avons utilisé les corpus *Multi-Domain Sentiment Dataset* (MDSD), collectés par (Blitzer *et al.*, 2007). Il s'agit de quatre corpus thématiques (*DVDs*, *kitchen*, *electronics* et *books*) contenant des critiques collectées sur le site internet Amazon. Chacun des corpus thématiques contient 1000 critiques positives et 1000 critiques négatives que nous utilisons pour la détection des marqueurs multi-polaires. Ces corpus contiennent également un certain nombre de critiques supplémentaires qui seront utilisées pour le test des expériences présentées dans les parties suivantes (de 3586 à 5945 selon le corpus). Les textes sont représentés en sacs de mots de bi-grammes et uni-grammes des formes fléchies des mots pleins. Leurs nombres d'occurrences sont pondérés par la taille du texte.
Pour le français, nous avons utilisé les corpus *JeuxVideo* et *AvoirAlire* issus du Défi Fouille de Textes 2007 (DEFT) (Grouin *et al.*, 2007). Ces corpus contiennent des critiques issues des sites *avoir-alire.com* et *jeuxvideo.com*. Elles sont réparties en trois classes, positif, neutre et négatif mais nous ne considérons ici que les classes positif et négatif. Comme le corpus *AvoirAlire* contient des critiques de différents domaines (films, musiques, livres, pièces de théâtre...), une séparation manuelle selon ces sous-domaines a été effectuée. Les critiques de ces corpus sont majoritairement étiquetées positif ou neutre. Seule la sous-partie *films* contient suffisamment de critiques négatives pour représenter un corpus d'apprentissage équilibré. Pour la détection des marqueurs multi-polaires, nous avons donc utilisé des corpus constitués de critiques sélectionnées au hasard dans la sous-partie *films* de *AvoirAlire* ainsi que dans *JeuxVideo* afin de constituer deux corpus thématiques équilibrés. Chacun contient 420 critiques positives et 420 critiques négatives. Le reste des critiques est utilisé pour le test lors des expériences des parties suivantes (293 textes pour *films*, 1446 pour *jeux vidéo*). Comme pour l'anglais, les textes sont représentés en sacs de mots pondérés de bi-grammes et uni-grammes des formes fléchies.

L'expérience se déroule de la manière suivante :

Détection des marqueurs multi-polaires
Pour l'extraction des marqueurs multi-polaires, nous utilisons les sous-parties annotées des domaines source et cible et réalisons le test du χ^2 comme décrit à la section 2.1. La sous-partie annotée du corpus cible n'est par contre pas utilisée pour entraîner le classifieur d'opinion. En effet, l'objectif de ce test est de valider que les mots multi-polaires ont un impact sur la détection d'opinion quand on change de domaine. Nous utilisons cette supervision pour extraire

les marqueurs multi-polaires, de façon à produire les meilleurs marqueurs effectifs. Dans un cadre réel d'adaptation au domaine, cette détection doit être faite de façon non supervisée, sans annotation dans le domaine cible.

Entraînement des classifieurs sur le corpus source modifié
Pour la classification automatique des textes en opinion positive/négative, nous avons utilisé un algorithme de boosting : *AdaBoost* dans son implémentation *BoosTexter* (Freund *et al.*, 1996; Schapire & Singer, 2000). Comme décrit à la partie 2.2, trois classifieurs d'opinion sont entraînés sur le corpus d'entraînement du domaine source : un premier classifieur de référence sans rien modifier, un classifieur entraîné en distinguant les marqueurs entre source et cible et un classifieur entraîné en supprimant tout simplement ces marqueurs de tous les corpus.

Classification du corpus cible modifié
Pour le test, nous utilisons la totalité des textes disponibles du domaine cible (la petite sous-partie ayant servie à la détection des marqueurs multi-polaires ainsi que tous les textes de test supplémentaires).

3.2 Exemples de marqueurs multi-polaires

Le tableau 1 présente quelques marqueurs détectés comme changeant de polarité entre deux domaines dans le corpus anglais MDSD. Pour chaque domaine, un mot a un score de positivité qui correspond à son nombre d'occurrences dans des critiques positives par rapport à son nombre total d'occurrences dans le domaine. Un score de 1 (resp. 0) signifie que dans ce domaine, le mot n'apparaît que dans des critiques positives (resp. négatives). Un écart de 0.5 est donc très significatif, faisant passer un mot de neutre à fortement polarisé.

	region	*I loved*	*worry*	*compare*	*return*
Domaine *electronics*	0.154	0.091	0.929	0.846	0.055
Domaine *books*	0.818	0.735	0.3	0.263	0.633

TABLE 1 – Pourcentage de présence de cinq exemples de marqueurs dans les critiques positives pour deux domaines. Le score va de 0 (très fortement négatif) à 1 (très fortement positif).

Nous avons ainsi en moyenne détecté 400 marqueurs multi-polaires sur l'anglais et 1000 sur le français. Ce décalage est vraisemblablement dû au fait que le vocabulaire français est, dans notre exemple, plus étendu que le vocabulaire anglais. Ceci s'explique d'une part parce que le corpus français considéré est de nature légèrement différente : les auteurs des textes étant des critiques de métier, ils ont vraisemblablement un vocabulaire plus riche que les auteurs des critiques du site Amazon. D'autre part, cette détection s'appuie sur les formes fléchies des mots, qui sont plus nombreuses en français du fait d'une morphologie plus riche. Notons que l'intégralité des marqueurs détectés selon cette méthode ne sera pas forcément utilisée par les classifieurs automatiques d'opinion. Il s'agit essentiellement d'indicateurs pour repérer d'éventuelles difficultés dans l'adaptation d'un domaine à un autre.

3.3 Évaluation

Les figures 1 et 2 présentent les résultats obtenus en exactitude (*accuracy*) respectivement pour le français et l'anglais. Ces résultats montrent que notre méthode donne de bons résultats sur le corpus français. En revanche, sur le corpus anglais, les résultats sont mitigés, avec des améliorations statistiquement significatives pour la moitié des paires testées et deux cas de détérioration. Il est cependant intéressant de noter que les meilleures améliorations sont observées pour les paires de corpus ayant le plus de difficulté de transfert (ceux dont l'exactitude du classifieur sans modification est déjà faible). Notre méthode étant purement statistique, elle ne fait pas intervenir, dans son fonctionnement théorique, des objets spécifiques à une langue. Par contre, elle s'appuie sur la segmentation en mots et les formes fléchies, ce qui peut être une piste pour expliquer les différences observées. Nous avons l'intention, dans des travaux futurs, de nous pencher de manière plus approfondie sur la sensibilité de notre méthode à la langue des textes.

52% des traits sélectionnés par BoosTexter sont des bi-grammes mais seul 42% des marqueurs multi-polaires utilisés le sont. Ainsi, en proportion, les marqueurs multi-polaires détectés sont plus souvent des uni-grammes même si la part de bi-grammes reste importante. De façon générale, on note qu'en moyenne, parmi les premiers mots choisis par BoosTexter en tant que classifieurs faibles (entre 700 et 800 selon les paires), 12 % se retrouvent dans notre liste de mots changeant de polarité en anglais et 10 % en français. Ainsi, près de 10 % des règles peuvent propager une erreur.
Il est également intéressant de noter que certains sens d'adaptation marchent mieux que d'autres. En effet, bien que pour

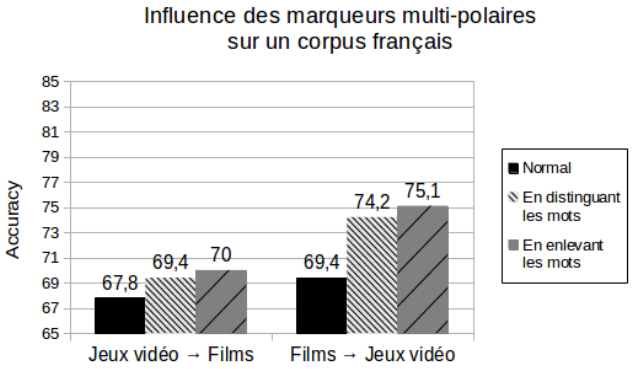

FIGURE 1 – *Accuracy* pour un classifieur entraîné sur un domaine source et testé sur un domaine cible en français (DEFT).

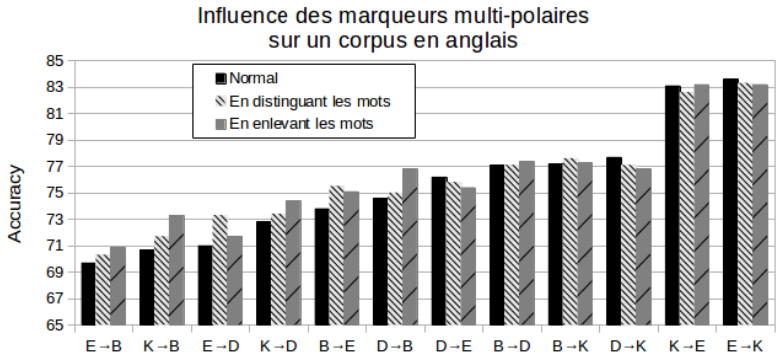

FIGURE 2 – *Accuracy* pour un classifieur entraîné sur un domaine source et testé sur un domaine cible en anglais (MDSD) ; D : DVD, B : books, E : electronics, K : kitchen.

une même paire de domaines les marqueurs changeant de polarité soient les mêmes dans un sens ou dans l'autre, ces marqueurs ne sont pas forcément utilisés comme traits dans les deux sens. Il est beaucoup plus fréquent que des marqueurs polaires deviennent neutres plutôt qu'ils passent de positif à négatif. Ainsi, si un marqueur est positif pour le domaine *films*, il sera appris comme trait positif par le classifieur. Et s'il est neutre pour le domaine *jeux vidéo*, cela provoquera une erreur de transfert. Mais dans l'autre sens, en s'entraînant sur *jeux vidéo*, rien ne sera appris pour ce trait puisqu'il est neutre. Ainsi, il n'y aura pas d'erreur de transfert bien que l'on perde de l'information. Par exemple, sur le corpus français, 92 marqueurs multi-polaires sont utilisés à l'origine dans le sens *films* vers *jeux vidéo* mais seulement 55 dans le sens *jeux vidéo* vers *films*. Aussi, il y a plus d'erreurs évitées dans un sens que dans l'autre.

Ainsi, cette prise en compte différenciée élémentaire des marqueurs changeant de polarité améliore la classification de l'opinion. Il est de plus vraisemblable qu'une pondération des marqueurs changeant de polarité, plutôt qu'une suppression complète, donne de meilleurs résultats (Choi & Cardie, 2009).

4 Utilisation des marqueurs multi-polaires pour des corpus multi-domaines

La partie précédente montre que les marqueurs multi-polaires sont utiles lors de l'adaptation d'un domaine source à un domaine cible. Nous nous plaçons à présent dans le cas où les corpus d'entraînement et de test sont chacun composés de plusieurs domaines de manière équivalente. Cela peut être le cas lorsqu'ils sont issus de la même source qui est elle-même multi-domaines, par exemple un blog abordant plusieurs sujets. Nous supposons dans cette section que la décomposition en domaines du corpus et l'attribution de chaque texte à un domaine sont connus.

4.1 Méthode

Avec un corpus multi-domaine, nous proposons une méthode de prise en compte des marqueurs multi-domaines en deux étapes : (1) la détection de marqueurs multi-polaires spécifiques à chaque domaine, (2) la construction de classifieurs d'opinions spécifiques à chaque domaine intégrant ces marqueurs.

4.1.1 Détection des marqueurs multi-polaires

Le processus de détection des marqueurs multi-polaires pour un corpus multi-domaine est présenté dans la figure 3. Le corpus d'entraînement est séparé en plusieurs sous-parties, chacune correspondant à un domaine particulier. Pour détecter les marqueurs multi-polaires, nous utilisons les étiquettes positives et négatives des données d'entraînement, comme décrit dans la section 2. Nous effectuons cette détection pour chaque sous-partie. A chaque fois, nous détectons les mots qui changent de polarité entre une sous-partie particulière du corpus d'entraînement et son complément (tous les autres textes). A la fin de cette procédure, nous avons plusieurs collections de marqueurs multi-polaires (une collection différente pour chaque sous-partie).

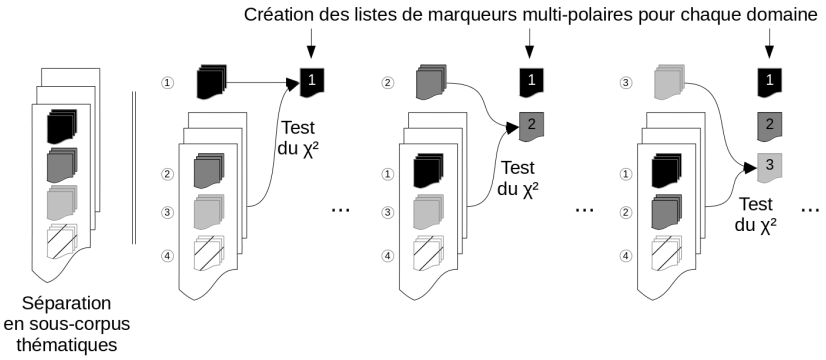

FIGURE 3 – Détection des marqueurs multi-polaires entre les sous-parties thématiques du corpus d'entraînement.

4.1.2 Différentiation des marqueurs multi-polaires

Nous créons un corpus d'entraînement différent pour chaque domaine par modification du corpus original en utilisant la liste de marqueurs multi-polaires associée à ce domaine. Pour cette expérience, nous avons testé uniquement la suppression des marqueurs multi-polaires. En effet, cette modification a donné globalement de meilleurs résultats dans nos précédentes expériences. Nous entraînons ensuite un classifieur sur ce corpus modifié et obtenons ainsi un classifieur spécifiquement adapté pour chaque domaine.

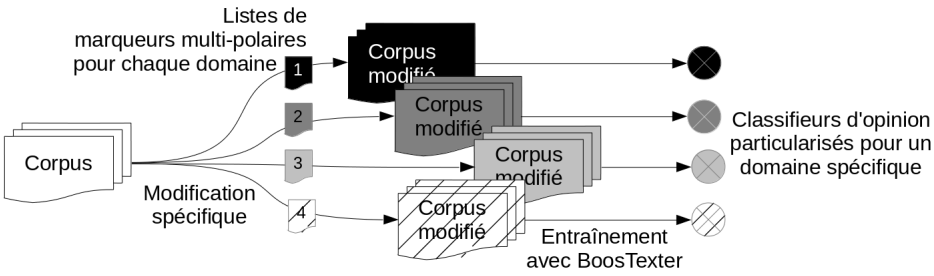

FIGURE 4 – Processus de création de plusieurs classifieurs thématiques en particularisant le corpus d'entraînement en enlevant les marqueurs multi-polaires des différentes listes associées à un domaine particulier.

Nous obtenons ainsi plusieurs classifieurs différents, chacun particularisé pour un domaine particulier (figure 4). Un texte du corpus de test est ensuite classifié en utilisant le classifieur propre à son domaine.

4.2 Évaluation des résultats

Nous avons effectué une évaluation de la méthode proposée avec le corpus *AvoirAlire* de DEFT dans son intégralité. Il y a donc 5 domaines : *livres* (757 textes), *bandes dessinées* (387 textes), *films* (1623 textes), *musique* (343 textes), *théâtre* (289 textes). Ils contiennent trois classes non équilibrées (55 % de textes positifs, 30 % de neutres et 15 % de négatifs). Pour l'anglais, nous avons de nouveau utilisé le corpus MDSD (8000 critiques annotées en positif/négatif réparties en quatre domaines : *DVDs*, *books*, *electronics* et *kitchen*).
Pour chaque corpus, nous avons réalisé une validation croisée. Le corpus est séparé aléatoirement en dix parties, neuf d'entre elles servant successivement de corpus d'entraînement et la dixième de corpus de test. Les résultats présentés sont les résultats moyens des dix expériences. Les textes sont toujours représentés en sacs de mots des uni- et bi-grammes des formes fléchies. La métrique d'évaluation utilisée lors de cette expérience est la F-mesure moyenne des classes positives et négatives.

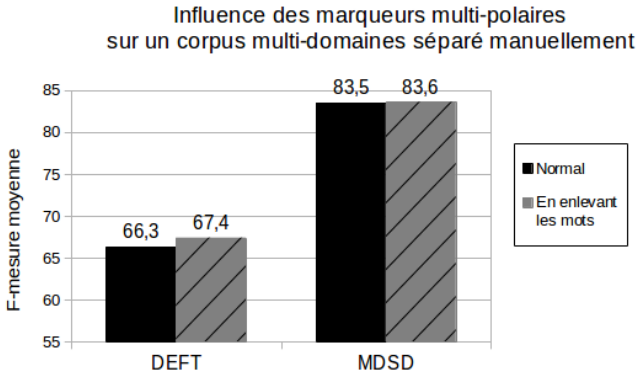

FIGURE 5 – Influence des marqueurs multi-polaires sur la performance d'un classifieur d'opinion au niveau du texte pour des corpus multi-domaines français (DEFT) et anglais (MDSD).

Les résultats, présentés à la figure 5, montrent que la prise en compte des marqueurs multi-polaires peut contribuer à améliorer la classification de l'opinion dans le cas d'un corpus contenant plusieurs domaines. Les améliorations potentielles sont cependant plus faibles que celles obtenues lors de l'adaptation d'un domaine à un autre (cf. section 3). En effet, les corpus d'entraînement et de test ont la même répartition de domaines. Notre méthode permet d'éviter des erreurs lorsqu'un mot a une certaine polarité dans tous les domaines sauf dans un où il a une polarité différente. Un apprentissage global assignera à ce mot la polarité dominante. Les erreurs ne se présenteront que dans la sous-partie du corpus de test associée avec le domaine isolé alors que, pour les autres parties du corpus de test, il n'y aura pas d'erreurs. Les erreurs que l'on peut éviter avec notre méthode sont donc moins nombreuses dans ce cas que lors de l'adaptation d'un domaine à un autre présentée dans la partie 3. Néanmoins, différencier les marqueurs multi-polaires n'est pas très difficile à mettre en place et se conjugue aisément avec les autres méthodes de classification de l'opinion en leur permettant d'éviter un certain nombre d'erreurs.

5 Utilisation des marqueurs multi-polaires pour des corpus en domaine ouvert

Dans la section précédente, nous avons fait l'hypothèse que la répartition des textes en différents domaines était connue. Or, ce n'est pas forcément le cas : certaines collections de textes contiennent des documents de différents domaines sans séparation ni indication explicite des domaines couverts. C'est par exemple le cas de corpus collectés automatiquement sur des médias particuliers, comme Twitter, qui présentent pourtant en général un grand intérêt pour des systèmes de veille d'opinion.

5.1 Méthode

La seule différence par rapport à la section précédente est l'absence d'étiquette de domaine pour les textes des corpus. Il est donc nécessaire de détecter automatiquement les différents domaines sous-jacents afin de séparer le corpus d'entraînement général en plusieurs corpus thématiques plus petits avant d'appliquer la méthode précédente. Ensuite, nous détectons les marqueurs multi-polaires et les intégrons afin de réaliser plusieurs classifieurs selon la méthode présentée dans la section 4. Néanmois, dans le cas des domaines ouverts, l'appartenance d'un texte à un domaine n'est pas une information binaire : on a en général un poids d'association entre un texte et un domaine. Pour chaque texte, les résultats des différents classifieurs spécifiques aux domaines doivent donc être fusionnés pour obtenir la classification finale de l'opinion.

5.1.1 Génération de domaines

Comme le corpus initial n'a pas d'étiquette de domaine, nous devons tout d'abord identifier les domaines sous-jacents et assigner chaque texte à un domaine. Nous avons utilisé dans ce but une méthode automatique de détection de thèmes (*Topic Models*) et, plus précisément, la méthode d'allocation de Dirichlet latente (LDA) (Blei *et al.*, 2003). Dans le cadre de la détection d'opinion, la méthode LDA a déjà été utilisée pour l'analyse de critiques focalisées sur un aspect, qui est proche de notre travail : dans (Titov & McDonald, 2008a,b), les auteurs introduisent un modèle fusionnant des *topics* locaux et globaux et utilisent les annotations manuelles des critiques afin d'améliorer l'identification des différents *topics*. D'autres travaux, tels que (Zhang *et al.*, 2013; Li *et al.*, 2010), combinent au modèle LDA des informations de sentiment ou bien des techniques de Naives Bayes afin de sortir du modèle en sac de mots.
Pour notre expérience, nous avons utilisé l'implémentation de la méthode LDA proposée dans Mallet (McCallum, 2002), qui utilise la méthode d'échantillonnage de Gibbs afin d'inférer la distribution utilisée pour la création des modèles de *topics*. Après avoir déterminé les *topics* à l'aide du corpus d'entraînement, chaque texte est représenté par un vecteur dont la taille est le nombre de *topics*, et dont chaque composante est la proportion de mots du texte qui appartient au *topic* associé à la dimension correspondante.
Le corpus d'entraînement est ensuite séparé en sous-parties, ou domaines, chacun d'entre eux associé avec l'un des *topics* sous-jacents détectés. Un texte est simplement associé au *topic* avec lequel il a le plus d'affinité. Par exemple, si sa proportion de mots appartenant à un *topic* est 55 %, il fera partie de la sous-partie du corpus associée au domaine correspondant.

5.1.2 Détection, différentiation et fusion

La détection des marqueurs multi-polaires ainsi que la différentiation du corpus d'entraînement en plusieurs corpus d'entraînement thématiques s'effectuent de la même façon qu'à la partie 4 en utilisant la partition en domaines induite par la méthode LDA. Nous obtenons ainsi plusieurs classifieurs thématiques, un par domaine.
La différence se situe lors de la classification des nouveaux textes. En effet, les textes du corpus de test n'ont pas d'étiquette de domaine. Nous devons tout d'abord déterminer leur profil de *topics* en utilisant le modèle de *topics* de la LDA. Ensuite, nous appliquons tous les classifieurs sur les nouveaux textes et obtenons plusieurs réponses différentes, une pour chaque classifieur spécifique au domaine. Nous fusionnons ces réponses en utilisant comme pondération les poids de leur profil de *topics*. Nous avons testé plusieurs stratégies de pondération pour cette fusion et la plus efficace a été de prendre l'exponentielle du score obtenu avec la LDA.

5.2 Évaluation des résultats

5.2.1 Description des corpus

Pour évaluer la méthode proposée pour les corpus en domaine ouvert, nous avons effectué tout d'abord une expérience sur les mêmes corpus français (DEFT) et anglais (MDSD) afin de pouvoir comparer avec l'expérience précédente utilisant une séparation en domaines explicite. De façon complémentaire, nous avons également utilisé le corpus anglais de tweets issu de la campagne d'évaluation SemEval 2013 pour la tâche 2 d'annotation de l'opinion (Wilson *et al.*, 2013). Ce dernier corpus est représentatif d'une collection de documents en domaine ouvert et permet de varier le type de textes sur lequel appliquer notre méthode. Les tweets sont nettoyés de leurs adresses internet, les émoticônes sont extraits et le nombre d'occurrences d'un type particulier d'émoticône (pleurs, rire, cœur...) est considéré comme un trait additionnel pour le

classifieur. Nous avons sélectionné au hasard une sous-partie équilibrée de ce corpus (1633 de chaque classe). Pour ce corpus uniquement, nous avons lemmatisé les mots du texte. En effet, les tweets étant de très courts textes, les formes fléchies ont peu d'occurrences. Comme pour les autres corpus, nous utilisons un sac de mots des uni- et bi-grammes.

La figure 6 montre que l'utilisation d'une partition automatique avec la LDA n'a pas modifié le comportement que l'on obtenait en utilisant une partition manuelle. Nous obtenons toujours une amélioration modeste sur le corpus français et des résultats similaires sur le corpus MDSD. L'utilisation d'une séparation automatique en domaines par LDA peut donc remédier à l'absence d'étiquette de domaine sans perte de performance.

5.2.2 Discussion des résultats

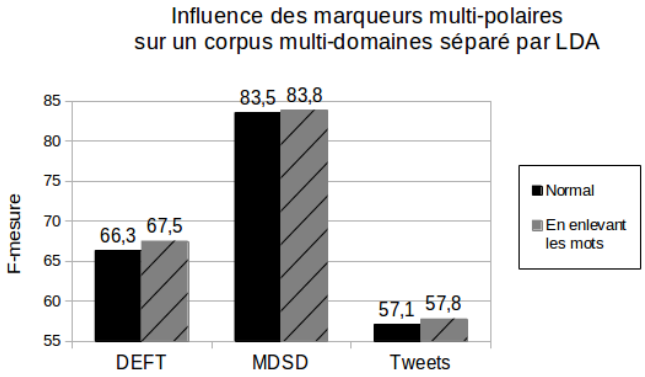

FIGURE 6 – Influence des marqueurs multi-polaires sur la performance d'un classifieur d'opinion au niveau du texte pour des corpus en domaine ouvert français (DEFT) et anglais (MDSD et Tweets). La séparation en sous-domaines thématiques a été effectuée par LDA.

Pour ce qui est du corpus de tweets, nous obtenons une très faible amélioration (+0.7 %) qui est néanmoins significative (selon un test de significativité par randomisation). Ce résultat doit être mis en relation avec le petit nombre de marqueurs multi-polaires détectés (en moyenne, 36 par domaine). Nous pensons que la taille du corpus, combinée aux 144 caractères des tweets, est trop petite pour que le test du χ_2 détecte beaucoup de marqueurs avec suffisamment de confiance. Pour comparaison, dans notre expérience sur les critiques en anglais, nous avions détecté 400 marqueurs multi-polaires par domaine. Nous nous sommes demandé si, pour ce corpus, des domaines plus focalisés sur un seul sujet pouvaient contrebalancer l'effet du manque de données.

Nous avons donc réalisé une seconde subdivision du corpus de tweets. Cette fois, un tweet n'est pris en compte que si plus de 75 % de ses mots appartiennent au même *topic*. Ainsi, un tweet dont seulement 55 % des mots appartiennent à un certain *topic* ne sera pas retenu. Dans cette version, les sous-parties du corpus d'entraînement obtenues sont plus focalisées sur un seul et même *topic*. En retour, elles contiennent moins de tweets et donc moins de données d'entraînement.

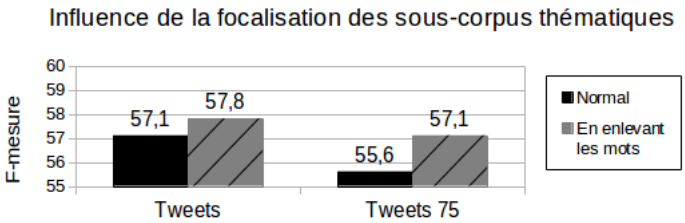

FIGURE 7 – Deux corpus d'entraînement différents sont utilisés. *Tweet* contient l'intégralité des tweets tandis que *Tweets75* contient uniquement ceux qui sont focalisés sur un seul *topic*.

La figure 7 montre le résultat de ces différentes intégrations des mots multi-polaires en utilisant deux corpus d'entraînement initiaux différents : avec l'ensemble des tweets ou avec uniquement les tweets les plus focalisés sur un *topic*. On remarque que pour l'expérience avec seulement les tweets les plus focalisés, l'amélioration est plus sensible (+1,46 % contre +0,70 %) bien que la valeur absolue du score reste inférieure en raison de la taille bien plus petite du corpus d'entraînement.

6 Conclusion

Dans cet article, nous avons étudié la notion de marqueurs multi-polaires d'opinion. Ce sont des mots ou groupes de mots qui sont indicateurs d'une certaine polarité d'opinion au niveau du texte en fonction du type d'objet dont le texte parle, ou domaine. Ces marqueurs multi-polaires sont de différents types linguistiques. Nous en avons proposé une première classification qui est actuellement en cours d'évaluation.
Nous avons testé l'apport de la prise en compte de ces marqueurs multipolaires pour la tâche de classification de l'opinion au niveau du texte lors de l'adaptation d'un domaine source à un domaine cible. Pour les corpus français étudiés, notre méthode présente une bonne amélioration de l'exactitude, allant jusqu'à +5,7 %. Pour les corpus anglais en revanche, il existe deux cas sur 12 pour lesquels la prise en compte de ces marqueurs multi-polaires dégrade significativement les performances. A l'inverse, dans 7 cas sur 12, notre méthode améliore significativement les résultats. Ces améliorations sont plus sensibles pour les paires de domaines pour lesquelles le transfert est le plus difficile.
Nous nous sommes également intéressés à l'apport possible des marqueurs multi-polaires pour la classification de l'opinion dans un corpus comportant plusieurs domaines. Nous particularisons le corpus d'entraînement pour chaque domaine et obtenons plusieurs classifieurs. Nos expériences montrent un gain moyen de +1,2 % de F-mesure pour le français. Le corpus de critiques en anglais n'obtient malheureusement pas d'augmentation significative. De plus, en l'absence de séparation explicite en domaines, le recours à un modèle de *topics* calculé par LDA ainsi qu'une fusion de classifieurs permettent d'obtenir les mêmes résultats. Nous avons également testé cette approche sur un corpus de tweets en anglais et nous trouvons une petite amélioration qui reste toutefois significative. Pour ce corpus contenant de très courts textes, nous avons montré que notre méthode est plus efficace lorsque les tweets composants le corpus d'entraînement sont focalisés précisément sur un seul domaine (+1,46 % de F-mesure).
Pour la suite de nos travaux, nous allons rechercher les phénomènes linguistiques qui influent sur la différence de performance entre français et anglais. Pour cela, nous allons poursuivre l'évaluation manuelle des marqueurs multi-polaires extraits automatiquement. Nous avons également l'intention de poursuivre nos expériences sur la façon de détecter ces marqueurs multi-polaires en utilisant le moins possible d'annotations dans le domaine cible. Une approche possible est de caractériser le comportement des mots candidats par rapport à des mots pivots de polarité stable et connue. Cela permettra de bénéficier de l'apport des marqueurs multi-polaires pour l'adaptation au domaine dans un cadre totalement non supervisé.

Références

AKKAYA C., WIEBE J. & MIHALCEA R. (2009a). Subjectivity word sense disambiguation. In *EMNLP*, p. 190–199, Singapore : Association for Computational Linguistics.

AKKAYA C., WIEBE J. & MIHALCEA R. (2009b). Subjectivity word sense disambiguation. In *EMNLP*.

BEN-DAVID S., BLITZER J., CRAMMER K. & PEREIRA F. (2007). Analysis of representations for domain adaptation. *Advances in neural information processing systems*, **19**, 137.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, **3**, 993–1022.

BLITZER J., DREDZE M. & PEREIRA F. (2007). Biographies, bollywood, boom-boxes and blenders : Domain adaptation for sentiment classification. In *ACL*.

CHOI Y. & CARDIE C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *EMNLP*.

DING X., LIU B. & YU P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of WSDM*, p. 231–240 : ACM.

FAHRNI A. & KLENNER M. (2008). Old wine or warm beer : Target-specific sentiment analysis of adjectives. In *Symposion on Affective Language in Human and Machine, AISB Convention*.

FREUND Y., SCHAPIRE R. E. *et al.* (1996). Experiments with a new boosting algorithm. In *ICML*, volume 96, p. 148–156.

GANAPATHIBHOTLA M. & LIU B. (2008). Mining opinions in comparative sentences. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, p. 241–248 : ACL.

GLOROT X., BORDES A. & BENGIO Y. (2011). Domain adaptation for large-scale sentiment classification : A deep learning approach. In *ICML*.

GROUIN C., BERTHELIN J.-B., EL AYARI S., HEITZ T., HURAULT-PLANTET M., JARDINO M., KHALIS Z. & LASTES M. (2007). Présentation de deft'07 (défi fouille de textes). *Actes du troisième DÉfi Fouille de Textes*, p.3̃.

HATZIVASSILOGLOU V. & MCKEOWN K. (1997). Predicting the semantic orientation of adjectives. In *EACL*, p. 174–181 : Association for Computational Linguistics.

HUANG F. & YATES A. (2012). Biased representation learning for domain adaptation. In *EMNLP*, p. 1313–1323, Jeju Island, Korea : Association for Computational Linguistics.

LI F., HUANG M. & ZHU X. (2010). Sentiment analysis with global topics and local dependency. In *Proceedings of the 24nd AAAI Conference on Artificial Intelligence (AAAI-10)*.

MCCALLUM A. K. (2002). Mallet : A machine learning for language toolkit.

NAVIGLI R. (2012). A quick tour of word sense disambiguation, induction and related approaches. *SOFSEM 2012 : Theory and Practice of Computer Science*, p. 115–129.

PAN S., NI X., SUN J., YANG Q. & CHEN Z. (2010). Cross-domain sentiment classification via spectral feature alignment. In *WWW*, p. 751–760 : ACM.

QUIRK R. & CRYSTAL D. (1985). *A comprehensive grammar of the English language*, volume 6. Cambridge Univ Press.

SCHAPIRE R. & SINGER Y. (2000). Boostexter : A boosting-based system for text categorization. *Machine learning*, **39**(2), 135–168.

SU F. & MARKERT K. (2008). From words to senses : a case study of subjectivity recognition. In *International Conference on Computational Linguistics*.

TAKAMURA H., INUI T. & OKUMURA M. (2006). Latent variable models for semantic orientations of phrases. In *EACL*.

TAKAMURA H., INUI T. & OKUMURA M. (2007). Extracting semantic orientations of phrases from dictionary. In *HLT-NAACL*, p. 292–299.

TITOV I. & MCDONALD R. (2008a). A joint model of text and aspect ratings for sentiment summarization. In *ACL*.

TITOV I. & MCDONALD R. (2008b). Modeling online reviews with multi-grain topic models. In *WWW*.

TURNEY P. & LITTMAN M. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Arxiv preprint cs/0212012*.

WIEBE J. & MIHALCEA R. (2006). Word sense and subjectivity. In *21st International Conference on Computational Linguistics and 44th annual meeting of the Association for Computational Linguistics*, p. 1065–1072 : ACL.

WIEBE J., WILSON T. & CARDIE C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, **39**(2-3), 165–210.

WILSON T., KOZAREVA Z., NAKOV P., RITTER A., ROSENTHAL S. & STOYANOV V. (2013). Semeval-2013 task 2 : Sentiment analysis in twitter. In *7th International Workshop on Semantic Evaluation*.

WILSON T., WIEBE J. & HOFFMANN P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP*.

WILSON T., WIEBE J. & HOFFMANN P. (2009). Recognizing contextual polarity : An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, **35**, 339–433.

WILSON T. A. (2008). *Fine-grained subjectivity and sentiment analysis : recognizing the intensity, polarity, and attitudes of private states*. ProQuest.

WU Y. & JIN P. (2010). Semeval-2010 task 18 : Disambiguating sentiment ambiguous adjectives. In *5th International Workshop on Semantic Evaluation*, p. 81–85.

YOSHIDA Y., HIRAO T., IWATA T., NAGATA M. & MATSUMOTO Y. (2011). Transfer learning for multiple-domain sentiment analysis - identifying domain dependent/independent word polaritys. In *AAAI*.

ZHANG Y., JI D.-H., SU Y. & WU H. (2013). Joint naïve bayes and lda for unsupervised sentiment analysis. In *PAKDD (1)*, p. 402–413.