Représentations et régularisations pour la classification de sentiments

Abdelhalim Rafrafi — Vincent Guigue — Patrick Gallinari

Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie, Paris 6
{abdelhalim.rafrafi, vincent.guigue, patrick.gallinari}@lip6.fr

RÉSUMÉ. *Les forums, les blogs et les recommandations sur les sites de vente en ligne constituent une source de données d'un nouveau genre présentant de forts enjeux économiques et scientifiques. L'exploitation de ces données permet de prédire efficacement les ventes de jeux vidéos et les entrées de cinéma. Le but de la fouille d'opinion est également d'affiner les profils d'utilisateurs et d'utiliser les sources ouvertes du web pour effectuer des sondages. Les algorithmes classiques de classification de documents ne fonctionnent pas de manière optimale sur ces données, ce qui explique la dynamique de recherche actuelle sur le sujet. Nous comparons dans cet article différents descripteurs textuels sur la tâche de classification supervisée de polarité et nous montrons l'intérêt des descripteurs complexes (N-grammes, sous-séquences) par rapport aux unigrammes. Ces représentations riches aboutissent à une très grande dimensionnalité qui pose problème lors de l'apprentissage: nous proposons une nouvelle méthode de régularisation basée sur la pénalisation des termes fréquents qui permet d'exploiter efficacement de tels espaces. Nous montrons l'intérêt de cette approche sur les données Amazon et Movie Reviews.*

ABSTRACT. *As web 2.0 is spreading, users get used to give their opinion on forums, blogs and e-commerce websites. This is a valuable piece of information for many applications such as consumer modeling, sales prediction or opinion survey. According to the literature, the efficiency of opinion mining tools will mainly relies on the ability of discriminating texts that express positive sentiments from texts that express negative ones. Previous experiments show that this task is difficult. We compare various classical descriptors and point out the interest of large representations of texts (N-grams, sub-sequences) for this task. The dimensionality of the data causes problems during the learning step: we demonstrate the inefficiency of the classical regularization framework as well as the interest of penalizing frequent terms. We demonstrate the efficiency of our approach on classical Movie Reviews and Amazon data-sets.*

MOTS-CLÉS : *Classification de sentiments, régularisation, pénalisation des termes fréquents*

KEYWORDS: *Sentiment classification, regularization*

CORIA 2012, pp. 285–300, Bordeaux, 21-23 mars 2012

1. Introduction

Les utilisateurs du Web sont de plus en plus habitués à laisser leurs opinions sur les blogs, les forums et les sites de e-commerce. La fouille de ces opinions est un domaine présentant de forts enjeux économiques et scientifiques. Il est en plein essor depuis une dizaine d'années. Plusieurs corpus ont été rendues publiques (par exemple Amazon (Blitzer *et al.,* 2007) ou *Movie Reviews* (Pang *et al.,* 2002)). L'état de l'art de (Pang *et al.,* 2008) fait autorité dans le domaine et pointe les difficultés liées à la fouille d'opinion (FO). (Pang *et al.,* 2008) détaillent les différentes tâches de la FO, depuis la construction d'une taxonomie des émotions jusqu'à la quantification émotionnelle. Ils montrent qu'une tâche se distingue comme essentielle pour beaucoup d'autres : la classification de polarité, c'est-à-dire le fait de détecter le sentiment positif ou négatif d'un texte. Les multiples expériences sur la classification de sentiments montrent la difficulté de la tâche[1] et insistent sur les améliorations encore possibles : l'extraction des caractéristiques discriminantes est beaucoup plus difficile que pour la classification thématique, l'information subjective est moins accessible, elle ne dépend pas directement du champ lexical.

Plusieurs études se sont focalisées sur la construction de caractéristiques discriminantes pour la polarité (Das *et al.,* 2001, Pang *et al.,* 2004, Matsumoto *et al.,* 2005) en utilisant le codage de la négation, l'analyse syntaxique (*Part of speech*) ou grammaticale. Cependant, (Pang *et al.,* 2008) concluent qu'il est difficile de tirer un avantage significatif de ces enrichissements.

Dans cet article, nous montrons que les représentations classiques en unigrammes ne sont pas adaptées à la détection de sentiments. Premièrement, nous illustrons rapidement les limites des approches unigrammes sur quelques exemples de classification de polarité. Deuxièmement, nous démontrons l'existence d'un biais fréquentiel dans la sélection des termes subjectifs à l'aide de SentiWordNet (Esuli *et al.,* 2006) : les mots fréquents sont sur-représentés par rapport à leur subjectivité tandis que les mots rares sont sous-utilisés voire éliminés. Enfin, nous insistons sur l'inefficacité des méthodes de régularisation standards (tant $\mathcal{L}_1$ que $\mathcal{L}_2$) pour extraire les termes discriminants. Ces régularisations renforcent même le biais fréquentiel et dégradent systématiquement les performances en test.

Le but de notre étude est de montrer que les approches promouvant les représentations simples (telles que les unigrammes) ou minimisant l'intérêt des représentations plus complexes sont en réalité bridées par le biais fréquentiel sus-mentionné. Nous proposons un nouveau cadre de régularisation pénalisant les termes fréquents qui permet de sélectionner efficacement les caractéristiques discriminantes pour la classification de sentiments. Cette approche permet d'exploiter efficacement les grands espaces de description comme les N-grammes ou les sous-séquences (Matsumoto *et al.,* 2005) et nous présentons des performances significativement au-dessus de l'état de l'art sur les corpus classiques Amazon et Movie Reviews.

---

1. c.f. la liste de références `http://www.cs.cornell.edu/people/pabo/movie-review-data/otherexperiments.html`

Nous décrivons les différents modèles utilisés dans la section 2 et nous présentons les données en section 3. Nous démontrons ensuite les limites des unigrammes et du cadre classique de la régularisation dans la section 4. Nous donnons enfin les résultats de nos expériences en section 5.

2. Notations, modèles et algorithmes d'apprentissage

Notre cadre est celui de la classification supervisée de sentiments dans les documents textuels en limitant les étiquettes possibles à deux catégories : les positives et les négatives (les documents neutres sont éliminés comme c'est le cas dans la majorité des expériences (Pang *et al.,* 2002, Blitzer *et al.,* 2007, Whitehead *et al.,* 2009)). $\mathbf{X}$, $\mathbf{Y}$ représentent respectivement un corpus de N documents textuels en sac de mots et les étiquettes associées. $\mathbf{x}_i$ est le document i dont le $j$ème terme est noté x_{ij} et l'étiquette $y_i \in \{+1, -1\}$. La taille du vocabulaire est notée V.

Nous utiliserons toujours un codage présentiel plutôt que fréquentiel comme c'est le cas dans la littérature (Pang *et al.,* 2008) : les performances ont été systématiquement plus intéressantes en utilisant ce codage, quel que soit le corpus sentimental utilisé. Sauf mention contraire, $\mathbf{x}_i \in \{0, 1\}^V$. Nous nous limitons à l'étude des classifieurs linéaires : $f(\mathbf{x}_i) = \sum_{j=0}^{V} x_{ij} w_j, \mathbf{w} \in \mathbb{R}^V$ où w_j est le poids associé au $i$ème terme du vocabulaire. Etant donné que la prédiction est associée au signe de f, la valeur $|w_j|$ peut être vue comme l'importance du terme j. Dans le problème de classification de sentiments $|w_j|$ sera associée à une mesure de subjectivité.

2.1. *Cadre classique d'apprentissage : SVM, LASSO, Spline et L_1-SVM*

L'apprentissage supervisé d'un classifieur linéaire revient la plupart du temps à optimiser la formulation générique suivante :

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} \left(C(\mathbf{X}, \mathbf{Y}) + \lambda \Omega(f) \right) \quad [1]$$

où C est une fonction de coût quantifiant les erreurs de l'estimateur f sur la base d'apprentissage tandis que Ω représente la régularisation ayant pour but de prévenir le sur-apprentissage. λ est le compromis de régularisation et $\mathbf{w}^\star$ désigne les paramètres de l'estimateur optimal.

Quatre formulations classiques sont issues de ce cadre générique. Elles correspondent aux combinaisons de deux fonctions de coût (charnière et moindres carrés) et de deux fonctions de régularisation (basées respectivement sur les normes L_1 et L_2 du vecteur de paramètres $\mathbf{w}$). Le tableau 1 synthétise ces formulations.

Ces modèles de base permettent d'illustrer des comportements différents en terme de sélection d'informations discriminantes. Ils couvrent une partie importante de la bibliographie sur l'apprentissage supervisé. La fonction coût charnière (*hinge loss*) est une approximation convexe assez fine de l'erreur de classification. Elle se focalise sur

		Régularisation	
		L_1 $\Omega^{(L_1)}(f) = \sum_{j=1}^{V} \lvert w_j \rvert$	L_2 $\Omega^{(L_2)}(f) = \sum_{j=1}^{V} w_j^2$
Coût	Charnière (*hinge*) : $C_h(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{N} (1 - y_i f(\mathbf{x}_i))_+$	L_1 SVM (Bradley *et al.,* 1998)	SVM (Boser *et al.,* 1992)
	Moindres carrés : $C_{ls}(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{N} (y_i - f(\mathbf{x}_i))^2$	LASSO (Tibshirani, 1996)	Spline (Tikhonov, 1963)

Tableau 1 – Les quatre formulations classiques (et leurs références) obtenues en combinant deux fonctions coûts avec deux fonctions de régularisation.

les documents ambigus proches de la frontière de décision. A l'inverse, les moindres carrés sont plus souvent utilisés en régression. Ils optimisent un critère de corrélation et se focalisent sur les moyennes des documents d'une classe pour construire la frontière de décision. La régularisation $\mathcal{L}_2$ permet généralement de lutter contre le sur-apprentissage en conservant de bonnes performances. Lorsque nous utilisons la descente de gradient pour optimiser le vecteur $\mathbf{w}$, la mise à jour est la suivante : $\mathbf{w} \leftarrow \mathbf{w} - 2\varepsilon\mathbf{w}$ (en tenant seulement compte de la régularisation) : un poids w_j non nul n'est donc jamais mis à zéro. A l'inverse, la régularisation $\mathcal{L}_1$ est parcimonieuse : durant la descente de gradient, la mise à jour est effectuée selon $\mathbf{w} \leftarrow \mathbf{w} - \varepsilon \text{ sign}(\mathbf{w})$. Si le signe de w_j change, le poids est mis à zéro (cf. (Friedman *et al.,* 2007)). Pour résumer, à chaque pas les w_j se déplacent vers 0 et les poids suffisamment faibles sont éliminés.

Il est important de remarquer que les deux régularisations fonctionnent de manière uniforme. Etudions en détail la mise à jour associée à l'un de ces classifieurs, les Splines. Pour un document $\mathbf{x}_i$, la dérivée partielle du coût régularisé par rapport à w_j est proportionnelle à $(y_i - f(\mathbf{x}_i))x_{ij} - \lambda w_j$: tous les poids sont diminués d'un facteur λ et les dimensions non nulles de $\mathbf{x}_i$ participent à une correction de l'estimateur (rappel : $x_{ij} \in \{0, 1\}$). En conséquence, les poids w_j des termes peu fréquents ($x_{ij} = 0$ la plupart du temps) sont fortement touchés par la régularisation tandis que les termes fréquents ($x_{ij} = 1$ plus souvent) participent beaucoup à la construction de $\mathbf{w}^\star$. Cette observation est valable pour toutes les formulations étudiées dans cette section.

2.2. *Régularisation sur la Fréquence des Termes (RFT)*

Notre étude sur le comportement de ces classifieurs (Sec. 4 et Fig. 1) montre que la sélection des termes n'est pas pertinente sur la tâche de détection de polarité, quelles que soient les fonctions de coût et de régularisation testées. Les termes fréquents (pertinents ou non) sont mis en avant par toutes les stratégies précédentes. Nous proposons donc une nouvelle formulation pénalisant la fréquence des termes afin de pouvoir sélectionner les caractéristiques discriminantes sans biais : plus un terme est fréquent

plus le poids w_j associé sera pénalisé lors de la régularisation. La nouvelle formulation est la suivante :

$$\Omega(f) = \sum_{j=1}^{V} \nu_j \Omega_j(f), \quad \nu_j = \frac{\#\{\mathbf{x}_i | x_{ij} \neq 0\}}{\#\mathbf{X}} \in [0,1] \qquad [2]$$

où $\Omega_j(f)$ décrit la composante de Ω relative au terme j. ν_j est la fréquence documentaire du terme j dans la base d'apprentissage.

Le calcul du gradient de cette nouvelle formulation est trivial puisque : $\frac{\partial \Omega(f)}{\partial w_j} = \nu_j \frac{\partial \Omega_j(f)}{\partial w_j}$. Par rapport à la section 2.1, il est clair que les termes moins fréquents seront moins pénalisés : leurs influences sur la solution va augmenter mécaniquement. A l'inverse, la réduction des w_j des termes fréquents sera amplifiée : seuls les termes fréquents et discriminants contribueront à la solution.

Notre formulation peut être vue comme une variante des *Confident Weighted Models* (Dredze *et al.,* 2010, Crammer *et al.,* 2009). Cependant, leur approche est adaptative et se focalise sur les termes présentant une forte variabilité dans leur contribution à la solution. Notre approche n'est pas adaptative (la pénalisation est fixe au cours de l'apprentissage) et nous mettons l'accent sur les termes rares et discriminants (apparaissant dans une classe mais pas dans l'autre).

3. Données et solveur

Nous utiliserons cinq jeux de données pour illustrer le fonctionnement de nos approches : quatre sous-corpus d'Amazon (*Books, Dvd, Electronics et Kitchen*) (Blitzer *et al.,* 2007) ainsi que le corpus *Movie Reviews* (Pang *et al.,* 2004). Cette dernière est composée de commentaires longs (utilisant un vocabulaire large) tandis que les corpus Amazon contiennent des revues plus courtes et un vocabulaire plus restreint. Les dimensions des corpus sont fournies dans le tableau 2. Le corpus *Movie Reviews* est particulièrement intéressante au regard de l'importante bibliographie qui l'entoure[2] : cela nous permet d'évaluer nos performances dans un environnement compétitif. Le corpus Amazon permet d'évaluer nos stratégies sur des domaines très différents, où les sentiments sont exprimés de manière spécifique. Les corpus *Books et DVD* restent relativement proches tandis que le corpus *Electronics* décrit des produits d'électronique grand public. La base *Kitchen* est la plus éloignée des précédentes, elle décrit les produits liés à la cuisine. Nous utiliserons le corpus Amazon en mono-domaine (cas d'usage classique) et en multi-domaines (apprentissage sur un sous-corpus et test sur un autre sous-corpus). Tous nos résultats seront comparés avec l'état de l'art (Blitzer *et al.,* 2007, Pan *et al.,* 2010).

2. `http://www.cs.cornell.edu/people/pabo/movie-review-data/otherexperiments.html`

3.1. *Caractéristiques descriptives*

Dans toutes les expériences nous utilisons trois descriptions existantes, générant des espaces de dimensions croissantes : les unigrammes (U), bigrammes (UB) et les sous-séquences (UBS) inspirées de (Matsumoto *et al.,* 2005).

Comme nous l'avons déjà dit, nous utiliserons toujours un codage présentiel qui donne systématiquement de meilleures performances que les codages fréquentiels. Ce type de codage fait l'objet d'un consensus en classification de sentiments (Pang *et al.,* 2008). Il est intéressant de constater qu'il peut être vu comme une première étape dans la pénalisation des termes fréquents : tous les termes qui apparaissent plusieurs fois dans un document sont codés de la même manière que ceux qui n'apparaissent qu'une fois.

Aucun traitement linguistique de type *stemming* ou lemmatisation n'est utilisé. Nous appliquons un filtre syntaxique (*Part-of-Speech*) conservant les étiquettes morpho-syntaxiques suivantes : JJ JJR JJS RB RBR RBS NN NNS VB VBD VBG VBN VBP VBZ MD. Ce qui correspond grossièrement aux noms, adjectifs et verbes. Les mots rares (n'apparaissant qu'une seule fois) sont remplacés par leur *Part-of-Speech*.

La notion de sous-séquence est inspirée de (Matsumoto *et al.,* 2005). Dans chaque phrase de chaque document, toutes les combinaisons de mots de longueur inférieure ou égale à trois sont calculées et stockées. Par exemple, dans la phrase A B C, nous extrayons les caractéristiques ABC, AB, AC, BC, A, B, C. Tous les motifs apparaissant plus de deux fois dans l'ensemble d'apprentissage sont conservés.

Bases	Tailles des corpus (N)	Long. moy. des revues	Vocabulaire (V)		
			Unigr.	U+Bigr.	U+B+SSéq.
Books	2000	240	10536	45750	78664
Dvd	2000	235	10392	48955	89313
Electronics	2000	154	5611	30101	49994
Kitchen	2000	133	5314	26156	40773
Movie Rev.	2000	745	26420	148765	308564

Tableau 2 – Descriptions des cinq corpus. La taille du vocabulaire dépend de la représentation choisie.

3.2. *Paramètres du solveur*

Dans toutes nos expériences, nous utilisons une descente de gradient mini-batch inspirée de (Bottou *et al.,* 2004). Il s'agit d'un compromis entre la descente de gradient stochastique qui est très efficace mais pose parfois des problèmes de stabilité et la descente de gradient classique qui est stable mais très lente. Notre implémentation est relativement simple et très efficace : les besoins en mémoire sont proportionnels à la taille du mini-batch, c'est-à-dire faibles et la méthode converge très rapidement.

Nous utilisons 500 itérations au maximum (une itération correspond à un passage sur les données) avec un mini-batch de taille 20 et un ε de 0.5. Nous utilisons une politique adaptative classique pour le pas de gradient : ε est multiplié par 99% à chaque itération. Enfin, dans la pratique nous n'effectuons jamais 500 itérations car nous avons introduit un critère d'arrêt supplémentaire : quand l'ensemble d'apprentissage est parfaitement discriminé, nous sortons de la phase d'apprentissage.

4. Biais fréquentiel et limites des unigrammes

Nous utilisons le corpus *Books* pour montrer les limites des approches classiques en classification de sentiments et l'intérêt de la régularisation sur la fréquence des termes (RFT). Tous les jeux de données se comportent de la même manière dans nos expériences.

4.1. *Fréquence vs subjectivité*

SentiWordNet (Esuli *et al.,* 2006) permet de mesurer la subjectivité d'un certain nombre de mots du point de vue de la perception humaine. Nous avons calculé pour chaque terme du dictionnaire que nous avons utilisé la subjectivité au sens de SentiWordNet[3] et nous disposons également de la fréquence des termes du dictionnaire dans notre corpus. Sur la Fig. 1 (a), nous montrons l'évolution de la subjectivité moyenne des termes en fonction de leur fréquence sur *Books*. Il apparaît clairement que les termes rares contiennent en moyenne autant de subjectivité que les termes fréquents. Nous sommes également capables de mesurer l'importance que notre classifieur attribue à un terme en étudiant la valeur absolue du poids w_j qui lui est associé (cf. Sec. 2). En effet, nous nous limitons à des modèles linéaires et les poids w_j pondèrent des caractéristiques binaires (0 ou 1) liée au codage présentiel : un w_j positif participe donc directement à rendre le score final positif (et réciproquement pour un poids w_j négatif). La courbe rouge ($+$) illustrant l'influence moyenne des termes en fonction de leur fréquence met en évidence la sur-représentation des termes fréquents par rapport aux termes rares : les mots rares sont (en moyenne) ignorés alors même que SentiWordNet démontre leur importance (courbe bleue, ligne pleine). L'usage de notre formulation RFT permet de rééquilibrer les influences pour se focaliser sur la subjectivité indépendamment de la fréquence : comme le montre la courbe verte ($*$, les termes rares voient leur influence croître et les termes fréquents sont plus pénalisés. La distribution des mots du corpus en fonction de leur fréquence (Fig. 1 (b)) est très déséquilibrée : les mots rares sont beaucoup plus nombreux que les mots fréquents. Les techniques usuelles de régularisation passent donc à côté de la majeur partie des informations subjectives du corpus.

3. SentiWordNet propose une quantification des sentiments positifs et négatifs présents dans 117000 termes, essentiellement des unigrammes et des bigrammes. Les auteurs proposent une formule pour en déduire un score de subjectivité compris entre 0 et 1. Les mots absents de SentiWordNet sont considérés comme purement objectifs.

SVM-RFT		SVM	
Nb. occ. (+)	Nb. occ. (−)	Nb. occ. (+)	Nb. occ. (−)
11.85	15.04	24.46	249.17

Tableau 3 – Nombre d'occurrences moyen des termes du top 100 (100 termes les plus influents en positif et négatif) pour les classifieurs SVM-RFT et SVM sur *Books*

Le tableau 3 confirme les résultats ci-dessus et montre que notre algorithme fonctionne comme prévu : les 100 termes les plus influents du SVM-RFT ont une fréquence bien moindre par rapport aux 100 termes les plus influents des SVM classiques. Le même phénomène est observé sur tous les modèles et tous les corpus.

Nous avons également calculé un score de subjectivité pour les classifieurs en mêlant la subjectivité SentiWordNet et les poids des classifeurs selon la formule suivante :

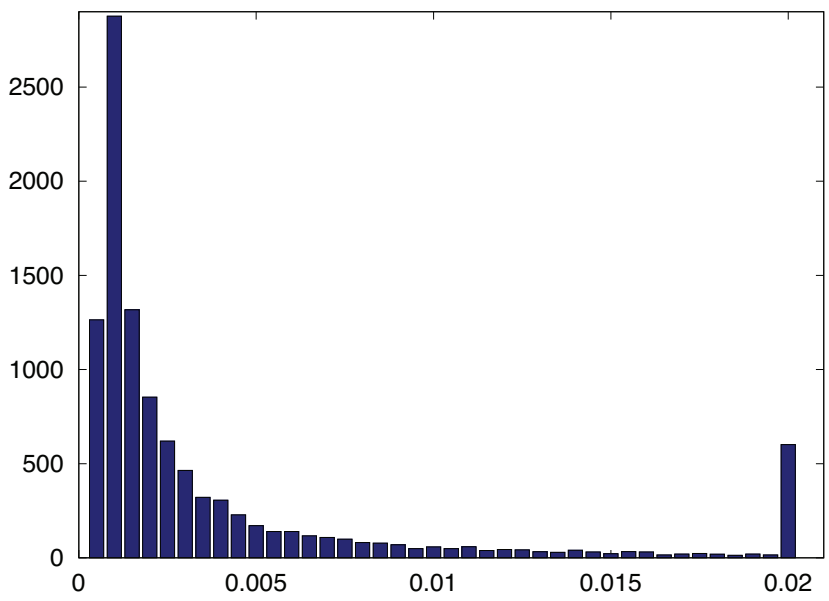
$$Subjectivité(\mathbf{w}) = \frac{\sum_{j=1}^{V} |w_j| subj_j}{\sum_{j=1}^{V} |w_j|}, \quad \forall j,\ subj_j \in [0,1]$$

Les SVM classiques et les SVM-RFT obtiennent respectivement des scores de 18.71% et 20.00% sur le corpus *Books*. Cette mesure permet encore une fois de valider l'intérêt de notre régularisation pour la détection de sentiments.

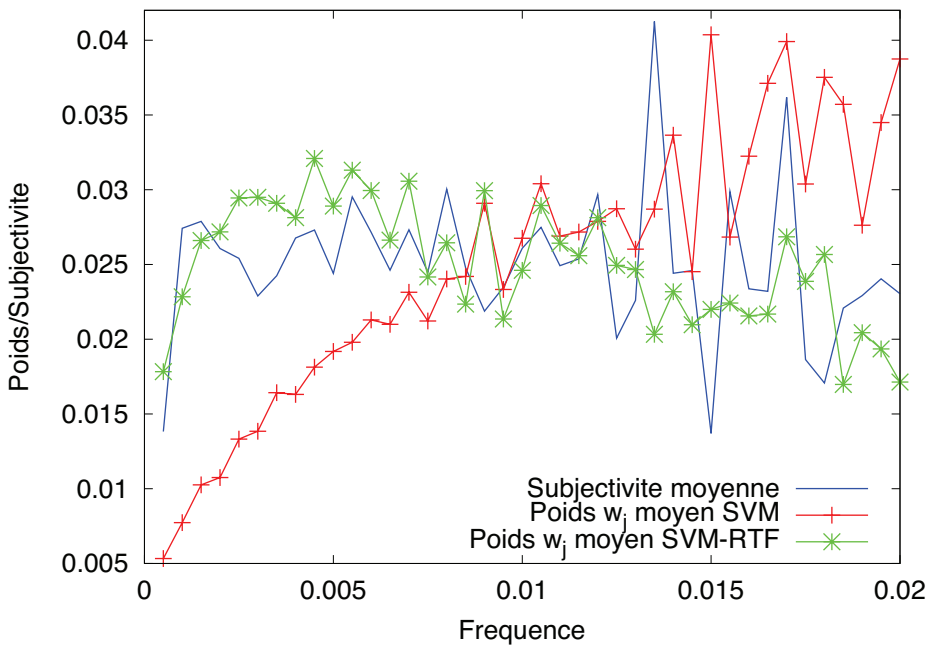
4.2. *Echec des régularisations classiques pour la sélection de termes discriminants*

La régularisation est l'approche la plus populaire pour contrôler la complexité et sélectionner les bonnes caractéristiques en apprentissage automatique. La plupart des algorithmes usuels d'apprentissage possède une version régularisée permettant de faire face au bruit des données et au problème de la dimensionnalité, lorsque le nombre de caractéristiques est important au regard du nombre d'échantillons. Ce dernier problème devient critique dès que nous utilisons des représentations avancées telles que les bigrammes ou les sous-séquences (cf. Table 2).

La figure 2 montre l'échec systématique des méthodes de régularisation classiques sur les bigrammes. Alors que le problème est particulièrement mal posé (45750 caractéristiques pour 2000 documents), la régularisation aboutit toujours à une baisse de performance en test et la procédure d'optimisation de λ (validation croisée) nous pousse à choisir une valeur nulle. Quel que soit le modèle de base choisi, nous sommes incapables d'éliminer le bruit présent dans les données : l'apprentissage est très rapide, l'erreur tendant vers 0 après quelques itérations seulement étant donné la dimension des données. Cette figure montre également la différence de comportement entre les régularisations $\mathcal{L}_1$ et $\mathcal{L}_2$: la première est plus sensible. En mettant plus de coefficients à zéros, elle provoque une rapide baisse de la dimension du problème et nous commençons rapidement à éliminer de l'information utile.



(a) Distribution de la fréquence des termes. Tous les termes apparaissant plus de 40 fois sont regroupés dans la dernière barre de l'histogramme. L'axe x donne la fréquence des termes, l'axe y le nombre de termes présentant cette fréquence.



(b) (Bleu, ligne pleine) subjectivité moyenne des termes par rapport à leur fréquence en se basant sur SentiWordNet. (Rouge+/Vert$*$) poids moyen des termes après l'apprentissage en fonction de leur fréquence (courbe rouge : SVM, $\lambda = 0$ et courbe verte : SVM-RFT, $\lambda = 0.001$).

Figure 1 – Les dictionnaires sont constitués majoritairement de mots apparaissant moins de 3 fois dans les corpus. La seconde courbe montre que les poids associés à ces termes rares sont sous-évalués par rapport à leurs subjectivités intrinsèques.

La Régularisation sur la Fréquence des Termes (RFT) permet d'effectuer une sélection des termes discriminants pour le problème de classification de sentiments : la performance augmente jusqu'à une valeur optimale de λ non nulle (elle chute ensuite de manière prévisible lorsque les données utiles sont éliminées). Plus l'espace descriptif est grand, plus le problème de la sélection est critique : la RFT est donc un outil important pour exploiter correctement les représentations riches. Cette expérience offre un nouveau point de vue sur la discussion concernant les espaces de représentation dans (Pang *et al.,* 2008) : alors qu'ils présentent le débat entre unigrammes et représentations riches comme ouvert au vu de la bibliographie, nous estimons que les représentations riches sont toujours plus intéressantes à condition d'avoir les outils pour les exploiter. Nous sommes également convaincus que le biais fréquentiel sus-mentionné est l'une des principales causes des performances décevantes parfois enregistrées avec les représentations riches.

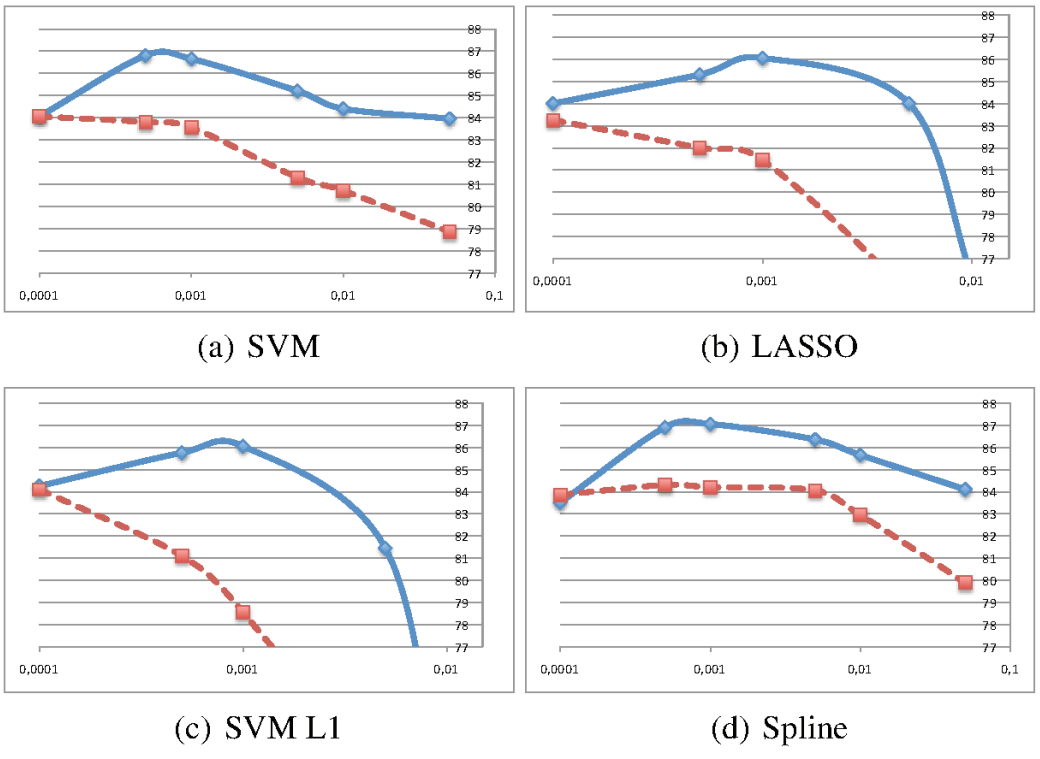


Figure 2 – Taux de reconnaissance des quatre modèles en fonction du paramètre λ sur le corpus *Books* en utilisant la représentation UB. Les figures permettent de comparer le fonctionnement de la régularisation classique (en rouge, pointillés) et la RFT (en bleu, pleine ligne). Tous les scores sont obtenus en validation croisée (5 ensembles).

4.3. *Limites qualitatives des unigrammes*

Les sacs de mots basiques ne sont pas bien adaptés à la classification de sentiments. Les linguistes insistent sur l'importance de la négation et de la structure des phrases dans l'expression des émotions. Les représentations enrichies permettent de mettre en évidence ce type de motifs (différencier le *good* du *not good* par exemple). Le tableau 4 illustre les limites des unigrammes pour la détection de sentiments : elle présente les

Unigrammes		Bigrammes	
top −	top +	top −	top +
disappointing	summer	not_recommend	a_must
useless	terrific	best_answer	I_sure
valuation	bible	save_your	really_enjoyed
outdated	displayed	too_much	read_from
shallow	editions	skip_this	excellent_book
poorly	refreshing	reference	wow
wasted	concise	disappointing	loved_this
unrealistic	profession	shallow	gift
norm	bike	unless_you	good_reference
hype	shines	way_too	enjoyed_this
incorrect	coping	was_looking	very_pleased
burn	blended	nothing_new	it_helped
boring	humorous	your_money	terrific
york	lighten	very_disappointing	great_ !
hated	amazed	first_trip	all_ages

Tableau 4 – Termes du top 15 (positif et négatif) pour les SVM-RFT sur le corpus *Books*.

quinze termes les plus représentatifs d'un classifieur SVM (en positif et en négatif) pour les unigrammes puis pour les bigrammes. Nous avons utilisé un SVM-RFT mais cela est anecdotique, le but est de montrer les limites des unigrammes : par exemple *recommend* n'apparaît pas dans les unigrammes mais est en tête des bigrammes dans la combinaison *not_recommend*.

En étudiant le tableau 4, plusieurs anomalies apparaissent dans les unigrammes : le mot le plus positif est *summer*. *Edition, profession, bike* suivent et nous imaginons bien qu'il s'agit de mots apparaissant en arrière plan des revues positives mais pas dans l'expression directe des émotions positives. A l'inverse, les bigrammes mettent en évidence systématiquement des expressions directes de sentiments : l'évaluation qualitative penche clairement du côté des représentations complexes. Il est également intéressant de noter la contribution de la ponctuation qui apparaît dans les bigrammes alors qu'elle n'émerge pas dans les unigrammes.

Les sous-séquences renforcent encore ce phénomène, notamment vis à vis du codage de la négation : quelle que soit la position du *not*, il sera associé à tous les mots de la phrase. (Pang *et al.,* 2008) (section 4.2.2) considère que le débat sur la représentation reste ouvert. Nous estimons que seules les représentations complexes sont efficaces.

5. Expériences

5.1. *Mono-domaine*

La Figure 3 propose une comparaison des quatre classifieurs de base avec leur version RFT. Tous les modèles se comportent de la même manière : la régularisation sur la fréquence des termes permet systématiquement de dépasser les modèles originaux[4]. La RFT permet d'extraire (ou de mettre en avant) des caractéristiques plus pertinentes quelle que soit la représentation.

Le tableau 5 compare les meilleures performances des modèles avec l'état de l'art issu de la bibliographie. Cette table montre l'intérêt de la combinaison entre une représentation riche (UB, UBS) et une méthode efficace de régularisation : cela nous permet de dépasser l'état de l'art sur tous les corpus considérés ici. Pour un classifieur donné, l'usage d'une représentation riche permet de gagner entre 1.3% et 3.4% en reconnaissance. La combinaison avec la RFT offre un gain de 4.4% à 6% par rapport au couple classique unigrammes+SVM. Cela montre que des gains importants sont encore réalisables sur ces données pourtant largement étudiées.

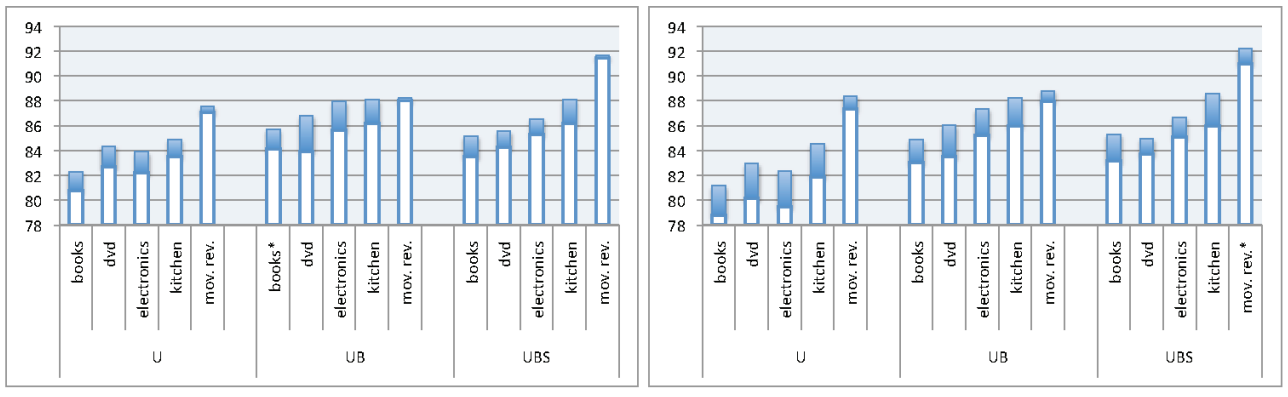
Alors que les meilleurs résultats sont obtenus avec les SVM dans le cadre classique, ce n'est plus le cas avec la RFT. Le meilleur modèle dépend alors du corpus : les Splines sont optimales pour Amazon (en moyenne sur les quatre sous-corpus) tandis que le LASSO est meilleur pour *Movie Reviews*. Globalement, les résultats des différents modèles sont proches mais il semble que le coût L_2 soit mieux adapté au cadre RFT.

L'exploitation de la représentation UBS pose encore problème : les résultats sur *Movie Reviews* montrent l'intérêt de l'approche. Cependant, la dimensionnalité de des données sous forme UBS est grande et le niveau de bruit important : les expériences sur Amazon montrent qu'il reste du travail pour exploiter pleinement cette représentation.

5.2. *Multi-domaines (Cross-domain)*

Les expériences multi-domaines consistent à apprendre sur un corpus (100% des échantillons) puis à tester sur un autre corpus. Nous avons utilisé les données Amazon pour réaliser ces tests en utilisant simplement la régularisation sur la fréquence des termes lors de l'apprentissage. Rappelons qu'Amazon est constitué de deux sous-corpus assez proches (*DVD* et *Books*), un sous-corpus assez général (*Electronics*) et un sous-corpus excentrique (*Kitchen*). Nous n'avons utilisé aucun transfert vers le domaine cible, aucun apprentissage non supervisé préalable et nous avons réutilisé directement les paramètres optimaux du mono-domaine sans chercher à les améliorer (la régularisation n'est pas plus forte ici) : le contexte de notre expérience est donc très

4. Pour certaines expériences de *Movie Reviews* (SVM L1 + UB, UBS et LASSO), les performances sont légèrement meilleures mais l'écart n'est pas significatif.



(a) SVM (b) LASSO

(c) SVM L1 (d) Spline

Figure 3 – Taux de reconnaissance des modèles pour trois représentations (U, UB et UBS). Chaque barre de l'histogramme montre la performance du modèle de base et de la variante RFT (NB : RFT est au dessus dans tous les cas). Les meilleurs résultats absolus sur chaque corpus sont marqués d'une étoile dans la légende.

différent (beaucoup plus simple et rapide) que ce qui se pratique habituellement dans la littérature.

Le tableau 6 rassemble les résultats de nos expériences croisées sur les sous-corpus d'Amazon. Comme dans les expériences mono-domaines, les meilleurs résultats sont distribués sur les différents algorithmes et les représentations UB et UBS. Les unigrammes proposent toujours des performances médiocres. UBS est un codage qui montre ici son intérêt (5 meilleurs résultats sur 12 expériences) mais aussi sa sensibilité : dès que les tâches s'éloignent (lorsque *Kitchen* est la cible par exemple), la performance chute.

L'usage de la régularisation RFT fonctionne donc correctement, ce qui n'était pas évident a priori : la régularisation fréquentielle augmente explicitement l'influence des petits coefficients, ce qui peut se traduire par du sur-apprentissage. La tâche de classification de sentiments multi-domaines requiert une généralisation particulièrement efficace. Nous montrons ici que la sélection de termes des modèles RFT est globalement pertinente pour ce problème. Cependant, la sensibilité des représentations UBS trouve sans doute son explication dans le sur-apprentissage de certaines itérations de la validation croisée.

	Meilleure RFT			Meilleure base			[1]	[2]	[3]	[4]
Modèle	Spline (Amazon), LASSO (M.R.)			SVM						
Caract.	U	UB	UBS	U	UB	UBS	U	UBS+	UB	UBS
Books	82.35	**85.5**	85.1	80.8	84.1	83.5	80.4	81.4	-	-
Dvd	84.25	**87.1**	85.6	82.7	84.0	84.3	82.4	82.55	-	-
Electr.	84.4	**88.2**	87.3	82.2	85.6	85.5	84.4	84.6	-	-
Kitchen	85.4	**88.3**	88.0	83.5	86.2	86.2	87.7	87.1	-	-
Mov. rev.	88.4	88.8	**92.2**	87.1	88	91.4	-	-	87.1	88.9

Tableau 5 – Meilleures performances obtenues en fonction des caractéristiques utilisées. Nos modèles sont comparés avec les modèles de bases ainsi qu'avec l'état de l'art. Toutes les performances sont mesurées en validation croisée (5 ensembles pour Amazon, 10 pour M.R. en accord avec la bibliographie).
Références : [1] (Blitzer *et al.*, 2007) [2] (Pan *et al.*, 2010) [3] (Pang *et al.*, 2004) [4] (Matsumoto *et al.*, 2005)
UBS+ correspond à des caractéristiques avancées issues de la langue naturelle.

Comparons nous maintenant avec l'état de l'art : (Blitzer *et al.,* 2007) utilise un critère d'information mutuelle (avec un jeu de données cibles non étiquetées) pour améliorer les performances sur les données cibles. Leur méthode nécessite donc plus de données et beaucoup plus de temps de calcul pour une performance optimale de 77.95% de reconnaissance (contre 77.75% pour nous). Nos résultats sont meilleurs dans 5 expériences sur 12 : les performances globales sont donc très proches malgré un contexte défavorable pour nous. (Pan *et al.,* 2010) améliore significativement les performances de (Blitzer *et al.,* 2007) en utilisant un nouveau critère pour le transfert. Ils obtiennent 78.68% de reconnaissance moyenne. Nos résultats ne restent optimaux que dans 3 expériences sur 12. Cependant leur méthode est particulièrement coûteuse.

6. Conclusion

Nous avons mis en évidence un biais fréquentiel dans la sélection des caractéristiques discriminantes par les méthodes usuelles d'apprentissage dans le cadre de la classification de sentiments. Une fois ce constat établi, nous proposons une nouvelle formulation pour la régularisation pénalisant la fréquence des termes. Ce nouveau cadre permet d'améliorer significativement les taux de reconnaissance des sentiments sur des corpus pourtant largement étudiés dans la littérature : en réalité, notre approche permet une sélection efficace des caractéristiques discriminantes et donc une exploitation optimisée des espaces de grandes dimensions.

Nous affirmons donc que le débat autour de l'intérêt des unigrammes pour la classification des émotions est biaisé : il est important d'utiliser des représentations plus riches pour capturer l'expression des sentiments, il suffit d'avoir des méthodes robustes pour traiter le problème d'optimisation résultant. Nous montrons les limites

	SVM			LASSO			L1 SVM			Spline		
Descr.	U	UB	UBS	U	UB	UBS	U	UB	UBS	U	UB	UBS
B → D	81.35	81.75	82.1	79.6	82.65	**83.5**	81.4	83.25	83.2	80.6	82.8	82.45
E → D	73.95	74.9	75.65	*68.3*	**77**	76.35	*72.95*	75.95	76.45	*72.1*	74.25	76.1
K → D	*73.15*	77.2	**77.5**	*70.3*	76.2	75.65	74.65	75.7	76	*71.45*	76.6	76.25
D → B	80.2	83.35	82.5	*78.35*	82.5	81.7	*78.95*	82.45	82.6	80.8	**83.6**	83
E → B	*68.95*	71.8	71.65	70.95	72.2	72.25	69.55	71.4	**72.9**	*68.55*	71.9	72.05
K → B	*69.6*	73.9	**74.1**	*67.6*	72.8	72.7	71.5	73.65	73.6	*68.35*	73.45	72.85
B → E	69.45	70.1	70.95	*68.1*	72	72.35	70.28	**72.4**	71.65	*67.85*	71.3	71.95
D → E	*69.9*	72.45	73.15	*68.3*	73.6	**74.8**	*70.8*	73.75	74.3	*70.7*	73.65	73.85
K → E	*81.5*	85.9	85.75	*79.05*	85.45	86.1	82.2	**86.15**	85.9	*80.65*	85.05	84.95
B → K	73.25	75.35	*71.1*	*70.5*	75.35	74.8	72.55	**75.8**	74.85	72.2	75.3	73.85
D → K	*72.1*	76.1	*73.55*	*71.75*	77	*73.4*	*72.85*	75.4	*73.05*	73.9	**77.65**	*73.4*
E → K	*81.9*	85.6	*78.9*	*78.55*	86.2	82.15	*81.5*	**86.3**	*81.85*	*81.25*	86.15	*79.65*
Moy.	74.61	77.37	76.41	72.61	**77.75**	77.15	74.93	77.68	77.20	74.03	77.64	76.70

Tableau 6 – Taux de reconnaissance sur les données Amazon dans le cadre multi-domaine. La première colonne décrit les sous-corpus utilisés (par exemple, B → D signifie que *Books* a été utilisée pour apprendre le modèle tandis que *DVD*, la cible, a été utilisée pour évaluer les performances). Les meilleures performances de chaque ligne sont mises en gras et les résultats médiocres ($\leq 95\%$ du meilleur) sont en italique. Les moyennes de performances sur les 12 expériences doivent être comparées à 78.65% (Pan *et al.*, 2010) et 77.95% (Blitzer *et al.*, 2007)

qualitatives des unigrammes sur un exemple frappant, le corpus *Books* du corpus Amazon.

Afin de rendre notre étude plus complète, nous avons étudié le comportement de quatre formulations usuelles en apprentissage sur cette tâche : les gains sont indépendants des coûts ou régularisations employés, le problème vient donc bien du biais fréquentiel mis en évidence précédemment.

Les perspectives autour de ce travail sont multiples : nous allons tester de nouvelles régularisations mixtes (type *Elastic Net*(Zou *et al.,* 2005) ou fused LASSO (Friedman *et al.,* 2007)). Nous essayons également d'améliorer la stabilité de nos approches pour exploiter au mieux l'espace UBS (qui est encore dépassé dans les expériences Amazon).

7. Remerciements

Ce travail est en partie financé par le projet DOXA de l'association Cap Digital, le pôle de compétitivité des contenus et services numériques.

8. Bibliographie

Blitzer J., Dredze M., Pereira F., « Biographies, Bollywood, Boom-boxes and Blenders : Domain Adaptation for Sentiment Classification », *ACL*, 2007.

Boser B., Guyon I., Vapnik V., « An training algorithm for optimal margin classifiers », *Workshop on Computational Learning Theory*, p. 144-152, 1992.

Bottou L., LeCun Y., « Large Scale Online Learning », *NIPS*, MIT Press, 2004.

Bradley P., Mangasarian O., « Feature selection via concave minimization and support vector machines », *ICML*, p. 82-90, 1998.

Crammer K., Kulesza A., Dredze M., « Adaptive Regularization Of Weight Vectors », *Advances in Neural Information Processing Systems*, 2009.

Das S., Chen M., « Yahoo ! for Amazon : Extracting Market Sentiment from Stock Message Boards », *Asia Pacific Finance Association Annual Conference*, 2001.

Dredze M., Kulesza A., Crammer K., « Multi-domain learning by confidence-weighted parameter combination », *Machine Learning Jour.*, vol. 79, n° 1–2, p. 123-149, 2010.

Esuli A., Sebastiani F., « SENTIWORDNET : A Publicly Available Lexical Resource for Opinion Mining », *Conference on Language Resources and Evaluation*, p. 417-422, 2006.

Friedman J., Hastie T., Hoefling H., Tibshirani R., « Pathwise Coordinate Optimization », *Annals of Applied Statistics*, vol. 1, n° 2, p. 302-332, 2007.

Matsumoto S., Takamura H., Okumura M., « Sentiment Classification using Word Sub-Sequences and Dependency Sub-Tree », *PAKDD*, 2005.

Pan S., Ni X., Sun J.-T., Yang Q., Chen Z., « Cross-Domain Sentiment Classification via Spectral Feature Alignment », *WWW*, 2010.

Pang B., Lee L., « A Sentimental Education : Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts », *ACL*, p. 271-278, 2004.

Pang B., Lee L., « Opinion mining and sentiment analysis », *Information Retrieval*, vol. 2, p. 1-135, 2008.

Pang B., Lee L., Vaithyanathan S., « Thumbs up ? : sentiment classification using machine learning techniques », *ACL-Empirical Methods in NLP*, vol. 10, p. 79-86, 2002.

Tibshirani R., « Regression shrinkage and selection via the lasso », *Journal Royal Statistics*, vol. 58, n° 1, p. 267-288, 1996.

Tikhonov A., « Regularization of incorrectly posed problems », *Soviet Math. Dokl.*, vol. 4, n° 6, p. 1624-1627, 1963.

Whitehead M., Yaeger L., « Building a General Purpose Cross-Domain Sentiment Mining Model », *IEEE World Congress on Computer Science and Information Engineering*, p. 472-476, 2009.

Zou H., Hastie T., « Regularization and variable selection via the Elastic Net », *Journal of the Royal Statistical Society, Series B*, vol. 67, p. 301-320, 2005.