Classification de Sentiments Multi-Domaines & Passage à l'Echelle

Abdelhalim Rafrafi — Vincent Guigue — Patrick Gallinari

Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie, Paris 6 - 4 place Jussieu F-75252 PARIS cedex 05
{abdelhalim.rafrafi, vincent.guigue, patrick.gallinari}@lip6.fr

RÉSUMÉ. *La classification de sentiments multi-domaines est un problème complexe: en effet, les distributions de caractéristiques sont alors différentes dans les ensembles d'apprentissage et de test. Différentes propositions permettent de limiter la baisse de performance inhérente à ce cadre. Cependant, la classification de sentiments est une tâche particulière car le web participatif nous donne accès à une quasi-infinité de données étiquetées. Cela soulève de nouvelles questions: à partir de quel volume de données les distributions d'apprentissage et de test convergent elles? Quand est ce que l'intérêt des techniques de transfert disparait? Dans cet article, nous étudions le taux de reconnaissance en sentiments par rapport la taille des ensembles d'apprentissage.*

ABSTRACT. *Multi-domain sentiment classification is known to be a difficult task in the literature since the feature distributions are different on training and testing sets. Thus, different transfer learning techniques have been proposed to cope with the induced lack of performance in recent years. But, the sentiment classification task is a particular supervised task where the labeled data are almost infinite (on the web 2.0). As a consequence, a new question emerged: if we have enough labeled data, does the train distribution converge to the test distribution? When does the transfer learning benefit vanish? In this article, we study the sentiment classification accuracy wrt the learning set size on the Amazon dataset.*

MOTS-CLÉS : *Classification de Sentiments, Transfert Multi-Sources, Fouille d'Opinion*

KEYWORDS: *Multi-Source Domain Adaptation, Sentiment Classification, Opinion Mining*

1. Introduction

La fouille d'opinion s'est développée avec le web participatif (2.0) et les contenus utilisateurs. Comme le résument (Pang *et al.,* 2008) dans leur étude, deux problèmes principaux restent ouverts en classification de sentiments : la prise en compte de la structure dans l'expression des sentiments et le développement de modèles robustes en multi-domaines. Nous nous intéressons dans cet article au second problème en étudiant le comportement de techniques d'apprentissage automatique sur plusieurs domaines en classification de sentiments.

Les contenus utilisateurs (revues, commentaires...) sont souvent postés avec une note explicite (par exemple en étoiles) et les techniques d'apprentissage permettent d'exploiter ces données. L'efficacité de ces modèles en classification de sentiments a été démontré dès (Pang *et al.,* 2002). Plusieurs améliorations ont ensuite été proposée via des caractéristiques de haut niveau allant des n-grammes (Dave *et al.,* 2003) au codage des négations et de l'analyse d'arbre syntaxiques (Das *et al.,* 2001). Mais toutes ces approches sont dépendantes du domaine et construire des classifieurs plus robustes est vite devenu un enjeu majeur (Blitzer *et al.,* 2007) : la problématique est appelée *adaptation multi-domaine*. Pour tester la robustesse des solutions, le protocole consiste à apprendre sur un jeu de données (source) et tester les performances sur des données d'un autre domaine (cible). Les propositions se répartissent en trois grands ensembles : utiliser le cadre de la régularisation pour améliorer le pouvoir de généralisation des modèles(Dredze *et al.,* 2010a, Rafrafi *et al.,* 2012), apprendre un espace sémantique codant des informations complexes et générales qui facilitent le passage à des données inconnues (Liu *et al.,* 2007, Maas *et al.,* 2011) ou optimiser un modèle explicite de transfert qui minimise la distance entre les distributions source et cible (Blitzer *et al.,* 2007, Pan *et al.,* 2010).

Nous nous intéressons aux modèles d'adaptation à partir de sources multiples comme (Mansour *et al.,* 2008) : plusieurs bases d'apprentissage de différents domaines sont fusionnées pour apprendre un modèle qui est testé sur de nouvelles données. Nous montrons que dans ce cadre, il est possible d'obtenir systématiquement des performances meilleures qu'en intra-domaine (apprentissage et test réalisés sur la cible). A notre connaissance, il s'agit de la première étude aboutissant à cette conclusion. La taille des ensembles d'apprentissage est directement liée à la performance en reconnaissance de sentiments, pourtant, la plupart des études précédentes se focalisent sur de petits jeux de données qui ne permettent pas de modéliser correctement la distribution de mots caractéristiques des sentiments. Afin d'expliquer les bons résultats obtenus, nous analysons en détail l'évolution des performances de nos systèmes en fonction de la taille des corpus d'apprentissage.

En section 2, nous présentons les travaux connexes. Notre approche est détaillée en section 3. Enfin, toutes nos expériences sont décrites en section 4 : nous analysons successivement le cadre intra-domaine, les transferts mono-source puis multi-sources.

2. Travaux Connexes

Les algorithmes multi-domaines ont été largement étudiés en classification de textes, cependant, les applications en analyse de sentiments sont plus récentes (Blitzer *et al.,* 2007). L'hypothèse classique i.i.d. n'est pas valable dans le cas multi-domaines : chaque domaine est en effet associé à une distribution de mots caractéristique. Cela explique l'écart de performance par rapport aux systèmes mono-domaines. Les techniques d'adaptation consistent à améliorer la généralisation des algorithmes en utilisant différentes stratégies :

Régularisation : La plupart des classifieurs de sentiments sont linéaires et utilisent une représentation en sac de mots (combinée à un codage présentiel) (Pang *et al.,* 2008). Ces modèles sont généralement régularisés et des formulations spécifiques à la classification de sentiments ont été proposées (Rafrafi *et al.,* 2012, Dredze *et al.,* 2010a). (Daumé-III, 2007) propose un cadre particulier basé sur l'enrichissement de la représentation mais requérant des données étiquetées du domaine cible.

Alignement Explicite : Une autre possibilité consiste à chercher les points communs entre les distributions source et cible pour améliorer le transfert : (Blitzer *et al.,* 2007, Pan *et al.,* 2010) proposent de chercher les mots qui ont le même comportement dans les deux domaines. Une fois identifiés les mots *pivots*, ils utilisent des techniques de factorisations matricielles permettant de construire de nouvelles caractéristiques robustes aux transferts. Ces approches nécessitent des données non-étiquetées venant de la cible.

Apprentissage d'un Espace Sémantique Le concept d'espace sémantique consiste à utiliser de gros corpus de documents pour apprendre les positions des mots dans un espace métrique. Un tel système capture les synonymies et tire parti des mots de la cible même lorsqu'ils n'apparaissaient pas dans la source. Les premières approches se sont basées sur des dérivés de PLSA (Liu *et al.,* 2007), puis ces modèles ont été dépassés par ceux dérivés de LDA (Gerrish *et al.,* 2011). Les travaux plus récents utilisent beaucoup les réseaux de neurones : (Glorot *et al.,* 2011) se base sur les auto-encodeurs, (Bespalov *et al.,* 2011) sur les réseaux à convolutions et (Maas *et al.,* 2011) sur les réseaux récursifs.

Deux cadres distincts existent en classification de sentiments multi-domaines : le plus simple consiste à prendre un seul domaine pour l'apprentissage (mono-source) et un pour le test (cible) (Blitzer *et al.,* 2007). Cependant, l'intérêt pratique de cette approche est discutable : en situation réelle, face à une cible inconnue, toutes les données d'apprentissage utiles peuvent être utilisées sans restriction. (Mansour *et al.,* 2008) propose une étude théorique du gain associé à l'apprentissage multi-sources : les auteurs modélisent la base cible comme une mixture des différentes sources en quantifiant les contributions de chaque source. Dans (Dredze *et al.,* 2010b), les auteurs montrent que l'utilisation de plusieurs sources combinées est plus efficace qu'une source seule (même la meilleure). Mais leur système est complexe et un passage à l'échelle semble délicat : ils se limitent à l'étude des petits corpus Amazon (2000 documents).

3. Motivation et Approche

Les organisateurs de l'atelier (Blitzer *et al.,* 2011) constatent que malgré les avancées récentes en adaptation multi-domaines, la plupart des approches ne sont pas robustes. Nous estimons que cette conclusion est directement liée à la petite taille des ensembles d'apprentissage généralement utilisés et nous proposons une analyse de l'impact de la taille des corpus sur les performances multi-domaines des modèles appris.

Un article récent démontre clairement le lien entre les performances en reconnaissance de sentiments et la taille des ensembles d'apprentissage (Bespalov *et al.,* 2011) : les auteurs font varier la taille de l'ensemble d'apprentissage entre 1000 et 300k documents et constate une progression régulière du taux de reconnaissance en test (sans atteindre un plateau). Nous sommes convaincus que ce phénomène est critique pour l'adaptation de domaine. L'expérience précédente montre qu'un large ensemble d'apprentissage ne permet pas de modéliser complètement la distribution des mots pour la classification de sentiments alors que nous n'avons pas encore introduit le cadre multi-domaine. Nous pouvons estimer que les ensembles de 2k revues de (Blitzer *et al.,* 2007) ne permettent pas non plus d'obtenir un modèle fiable et général pour la distribution des mots. Les bonnes performances multi-domaines obtenues par (Glorot *et al.,* 2011) nous poussent également à utiliser des corpus plus larges (quelque soit le domaine de ces corpus et même s'ils ne sont pas étiquetés).

Pour toutes les expériences nous utilisons des SVM linéaires classiques. Nous souhaitons montrer qu'un modèle de base peut atteindre les performances de l'état de l'art en utilisant simplement plus de données. Pour cette raison, nous nous restreignons à une description classique des données (uni-grammes et bi-grammes) ainsi qu'au codage présentiel recommandé dans (Pang *et al.,* 2008).

4. Expériences

Nous proposons dans cette partie deux séries d'expériences. Dans un premier temps, nous étudions le corpus Amazon qui est largement utilisé dans la littérature (Blitzer *et al.,* 2007, Pan *et al.,* 2010, Dredze *et al.,* 2010a). Nous testons plusieurs cadres d'apprentissage : le cas i.i.d, dit intra-domaine, les cas d'adaptation mono-source et multi-sources.Nous montrons que les performances en transfert peuvent dépasser les performances du cas i.i.d. à condition d'avoir suffisamment de données en apprentissage.

4.1. *Données & Paramétrage des Expériences*

Les documents sont représentés en sacs de mots. Nous limitons le dictionnaire aux 5000 uni-grammes et bi-grammes les plus fréquents comme cela est généralement fait dans la littérature. Comme le préconnisent (Pang *et al.,* 2008), nous utilisons un

codage présentiel[1]. L'apprentissage est réalisé avec SVMLight (Joachims, 2002) en conservant tous les régalages par défaut (y compris la régularisation).

Le Amazon (Blitzer *et al.,* 2007) est divisé en 25 domaines et compte 100k documents en apprentissage. L'ensemble de test est fixé une fois pour toutes : il s'agit des parties de test des sous-corpus d'Amazon *Books*, *Dvd*, *Electronics* et *Kitchen*. Cela représente un total de 27847 documents (respectivement 10k, 10k, 4k et 4k). Les expériences montrent deux comportements types : d'une part *Books* et *Dvd* qui sont assez proches et d'autre part, *Electronics* et *Kitchen*, qui réagissent de la même manière aux différents tests.

En fonction des expériences, nous utilisons les ensembles d'apprentissage suivants :

– pour le cas i.i.d (intra-domaine), les parties d'apprentissage des domaines cibles sont utilisés (*Books*, *Dvd*, *Electronics* et *Kitchen*).

– Dans les expériences suivantes, nous utilisons les sous-domaines Amazon externes (différents de la cible). Amazon compte 25 domaines, il y a donc 24 domaines externes pour chaque cible. En adaptation mono et multi-sources, aucune donnée cible n'est utilisée lors de l'apprentissage.

4.2. *Adaptation Mono-Source*

A titre de référence, le tableau 1 donne les performances du SVM sur la tâche classique iid intra-domaine.

Books	91.1%	DVD	90.6%	Electronics	90.6%	Kitchen	91.7%

Tableau 1 – Taux de reconnaissance intra-domaines

Le transfert mono-source est le cadre le plus répandu en classification de sentiments (Blitzer *et al.,* 2007). Une fois le modèle appris sur un domaine source (un sous-corpus d'Amazon), le test est effectué sur *Books*, *Dvd*, *Electronics* et *Kitchen*. Le principal constat réside dans l'instabilité chronique des résultats présentés en Figure 1 : les performances varient énormément en fonction de la source et de la cible. Des sources comme *Health, Toys, Sports*, ou *Apparel* proposent des performances correctes sur toutes les cibles. D'autres sources comme *Office, Musical Instruments, Video*,ou *Foods* donnent de bon taux de reconnaissance sur les cibles *Books* et *Dvd* mais pas sur *Kitchen* et *Electronics*.

Notre analyse de la Figure 1 est la suivante :

– les cibles *Books* et *Dvd* ont des comportements similaires par rapport aux différentes sources. La même conclusion s'impose pour le couple *Kitchen, Electronics*.

1. Les caractéristiques sont binaires, nous n'utilisons ni tf ni tf-idf.

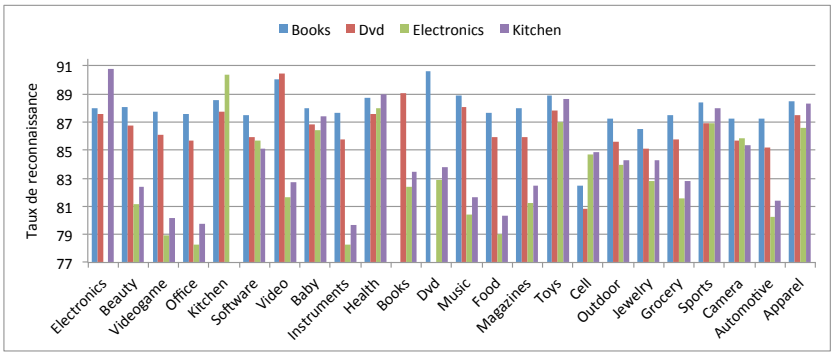


Figure 1 – Taux de reconnaissance sur les cibles *Books, Dvd, Electronics* et *Kitchen* en fonction des domaines sources. Les expériences intra-domaines ne sont pas reportées ici (une barre manque donc pour chaque domaine cible).

Dans les deux cas, les performances croisées à l'intérieur du couple (apprentissage sur l'un des membres et test sur l'autre) sont particulièrement élevées.

– D'une manière générale, les cibles *Books* et *Dvd* sont plus simple à classer que les cibles *Kitchen* et *Electronics*.

– Les performances varient beaucoup. Les taux de reconnaissance sur *Books* et *Dvd* vont de 80.8% à 90.6% en fonction de la source. C'est encore plus frappant pour *Kitchen, Electronics* (entre 78.3% et 90.7% de reconnaissance).

– Finalement ce cadre ne convient pas bien à l'adaptation : il requiert un oracle pour trouver la source optimale pour chaque cible.

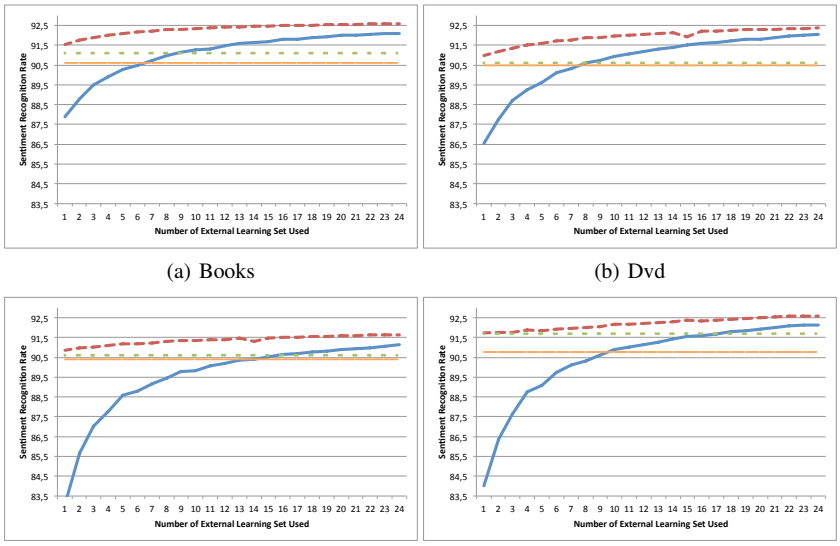
4.3. *Adaptation Multi-Sources*

Comme nous l'avons déjà dit, le cas multi-sources est un cadre plus réaliste que le précédent : nous cherchons à obtenir la meilleure performance sur une base inconnue à partir de tous les documents disponibles. Les quatre ensembles de test ne change pas mais *Books*, *Dvd*, *Electronics* et *Kitchen* ne sont plus utilisés en apprentissage. La figure 2 montre l'évolution des performances en fonction du nombre de sources externes utilisées, les taux de reconnaissance sont moyennés sur plusieurs expériences.

Nous tirons plusieurs conclusions de ces expériences :

– le taux de reconnaissance est toujours lié à la taille de l'ensemble d'apprentissage (et le plateau de performance n'est pas atteint).

– Le cas du transfert mono-source est particulièrement défavorable : l'écart au début de la courbe par rapport à la performance intra-domaine va de 3,3% à 7,7% mais il diminue rapidement dès que de nouvelles sources sont utilisées.



(a) Books (b) Dvd

(c) Electronics (d) Kitchen

Figure 2 – Taux de reconnaissance sur les 4 cibles en fonction du nombre de sources externes utilisées. Performance intra-domaine (pointillés verts), transfert mono-source+oracle (ligne fine orange), transfert multi-sources (bleu), intra-domaine+enrichissement (pointillés rouges).

– Le transfert multi-sources permet de dépasser systématiquement la performance mono-source+oracle (pour choisir la meilleure source). Nous reproduisons ici les résultats de (Dredze *et al.,* 2010b) mais avec un système nettement plus simple. Les gains en taux de reconnaissance vont de 0,7 à 1,6% comparé à l'oracle.

– Quelque soit la cible, la courbe multi-sources dépasse toujours la courbe intra-domaine (de 0,4 à 1,4%) : cette observation est nouvelle et à notre connaissance aucun article publié ne propose un tel résultat. Ces performances doivent être comparées notamment à (Glorot *et al.,* 2011) qui utilise le même jeu de données sans aboutir à un gain systématique.

– Par rapport aux performances intra-domaine+enrichissement, l'écart reste toujours inférieur à 0,5% : la complexité et le coût des modèles de transfert deviennent des freins importants face aux maigres perspectives de gain.

5. Conclusion

Notre première conclusion est que les performances intra-domaines sur Amazon peuvent être dépassées à condition d'avoir suffisamment de données étiquetées

(quelque soit la source d'où elles viennent). Ce résultat constitue la nouveauté et l'apport de cet article.

Remerciement : ce travail a été partiellement financé par le projet DIFAC (FUI 12).

6. Bibliographie

Bespalov D., Bai B., Qi Y., Shokoufandeh A., « Sentiment classification based on supervised latent n-gram analysis », *ACM CIKM*, p. 375-382, 2011.

Blitzer J., Cortes C., Rostamizadeh A., « Domain Adaptation Workshop : Theory and Application », *NIPS Workshop*, 2011.

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

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

Daumé-III H., « Frustratingly Easy Domain Adaptation », *ACL*, 2007.

Dave K., Lawrence S., Pennock D. M., « Mining the peanut gallery : opinion extraction and semantic classification of product reviews », *WWW*, ACM, p. 519-528, 2003.

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, 2010a.

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

Gerrish S., Blei D., « Predicting Legislative Roll Calls from Text », *ICML*, p. 489-496, 2011.

Glorot X., Bordes A., Bengio Y., « Domain Adaptation for Large-Scale Sentiment Classification : A Deep Learning Approach », *ICML*, 2011.

Joachims T., *Learning to Classify Text using Support Vector Machines*, Springer - Kluwer Academic Publishers, 2002.

Liu Y., Huang X., An A., Yu X., « ARSA : a sentiment-aware model for predicting sales performance using blogs », *ACM SIGIR*, 2007.

Maas A. L., Daly R. E., Pham P. T., Huang D., Ng A. Y., Potts C., « Learning Word Vectors for Sentiment Analysis », *Association for Computational Linguistics (ACL)*, 2011.

Mansour Y., Mohri M., Rostamizadeh A., « Domain Adaptation with Multiple Sources », *NIPS*, 2008.

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., « 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.

Rafrafi A., Guigue V., Gallinari P., « Coping with the Document Frequency Bias in Sentiment Classification », *AAAI ICWSM*, 2012.