Détection d'opinion : Apprenons les bons adjectifs !

Ali Harb, Gérard Dray, Michel Plantié, Pascal Poncelet, Mathieu Roche, François Trousset

Détection d'Opinion : Apprenons les bons Adjectifs !

Ali Harb[1,2], Gérard Dray[1], Michel Plantié[1], Pascal Poncelet[1], Mathieu Roche[2], François Trousset[1]

[1]*EMA-LGI2P, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France*
{ali.harb, gerard.dray, michel.plantie, pascal.poncelet, francois.trousset}@ema.fr
[2]*LIRMM Université Montpellier II CNRS 5506, 161 Rue Ada, F-34392 Montpellier*
{mathieu.roche, ali.harb}@lirmm.fr

RÉSUMÉ. *L'expression d'opinions sur Internet se développe de plus en plus. Récemment, de nouveaux travaux de recherche se sont intéressés à l'extraction automatique des opinions exprimées. Traditionnellement, l'extraction d'opinion dans les textes est basée sur la recherche des adjectifs pour cela, les méthodes existantes sont souvent basées sur des dictionnaires généraux. Malheureusement ce type d'approche trouve ses limites : pour certains domaines, des adjectifs peuvent être inexistants voire contradictoires. Dans cet article, nous proposons une nouvelle approche de création automatique de dictionnaire d'adjectifs qui intègre la connaissance du domaine. Les expériences menées sur des données réelles ont montré l'intérêt de notre approche.*

ABSTRACT. *Expressed opinions grows more and more on the Internet. Recently, extracting automatically such opinions becomes a topic addressed by new research work. Traditionally, detection of opinions is based on extracting adjectives. Existing methods are often based on general dictionaries. Unfortunately, main drawbacks of these approaches are that, for different domains, adjectives could not exist and could have an opposite meaning. In this paper we propose a new approach to the automatic creation of dictionary of adjectives that integrates the domain knowledge. The experiments conducted on real data show the usefulness of our approach.*

MOTS-CLÉS : *Fouille de Texte, Régles d'Association, Orientation Sémantique, Apprentissage, Classification.*

KEYWORDS: *Text Mining, Association Rules, Semantic Orientation, Machine Learning, Classification.*

Atelier FODOP'08, pages 59 à 66

1. Introduction

Avec le développement du Web, et surtout du Web 2.0, le nombre de documents décrivant des opinions devient de plus en plus important. Il devient ainsi possible de pouvoir donner son avis sur un produit ou sur un film. Récemment, les chercheurs de différentes communautés (Data Mining, Text Mining, Linguistique) se sont intéressés à l'extraction automatique de données d'opinions sur le Web. Traditionnellement, les approches de détection d'opinions cherchent à déterminer les caractéristiques d'opinions positives ou négatives à partir d'ensembles d'apprentissages. Des algorithmes de classification (se basant notamment sur différentes techniques linguistiques) sont alors utilisés pour classer automatiquement les documents extraits du Web (Plantié *et al.*, 2008). Dans cet article, nous nous intéressons plus particulièrement à l'étape d'acquisition du vocabulaire caractérisant une opinion positive ou négative d'un document. De manière à caractériser ces dernières, les principaux travaux de recherche considèrent que l'orientation sémantique d'une opinion est exprimée par l'intermédiaire des adjectifs (Turney, 2002, Taboada *et al.*, 2006, Voll *et al.*, 2007, Hatzivassiloglou *et al.*, 1997, Kamps *et al.*, 2004). Cependant, basées sur des dictionnaires existants ou sur des listes prédéfinies d'adjectifs, la plupart des approches se trouvent confrontées au problème suivant. Considérons, par exemple, les deux phrases "*The picture quality of this camera is high*" et "*The ceilings of the building are high*". Dans le cas de la première phrase (e.g. une opinion exprimée sur un film), l'adjectif *high* est positif. Par contre dans la seconde phrase (e.g. un document sur l'architecture), l'adjectif est neutre. Notre objectif dans cet article est de proposer une méthode d'apprentissage automatique des adjectifs correspondant à une opinion exprimée sur un domaine spécifique.

L'article est organisé de la manière suivante : la section 2 présente les principales techniques d'apprentissage d'opinion. Notre approche est décrite dans la section 3. La section 4 présente les expériences réalisées à partir de données réelles issues de blogs.

2. Travaux antérieurs

Comme nous l'avons vu précédemment, la plupart des approches utilisent l'adjectif comme principale source de contenu subjectif dans un document. En général, l'orientation sémantique d'un document correspond alors à l'effet combiné des adjectifs trouvés dans le document, en se fondant sur un dictionnaire d'adjectifs annotés (e.g. Inquirer (Stone *et al.*, 1966) qui contient 3596 mots étiquetés positifs ou négatifs ou HM (Hatzivassiloglou *et al.*, 1997) qui répertorie 1336 adjectifs). Plus récemment, de nouvelles approches ont enrichi l'apprentissage des adjectifs à l'aide de système comme WordNet (Miller, 1995). Dans ce cadre, il s'agit d'intégrer automatiquement les synonymes et les antonymes (Andreevskaia *et al.*, 2007) ; ou d'acquérir des mots porteurs d'opinions (Voll *et al.*, 2007, Hu *et al.*, 2004). La qualité du résultat final est fortement liée aux différents dictionnaires disponibles et surtout, elles ne sont pas capables de différencier les adjectifs en fonction du domaine spécifique visé (e.g. high). Pour pallier ce problème, les approches les plus récentes utilisent des méthodes statis-

tiques basées sur la co-occurrence d'adjectifs à partir d'un ensemble de mots germes. Le principe général dans ce cas est, à partir d'un ensemble d'adjectifs positifs et négatifs (e.g. *good*, *bad*), de rechercher les adjectifs situés à une certaine distance. L'hypothèse sous jacente, dans ce cas est la suivante : un adjectif positif apparaît plus fréquemment aux côtés des mots germes positifs, tandis que les adjectifs négatifs apparaissent le plus souvent aux côtés de mots germes négatifs. Même si ces approches sont efficaces, elles souffrent des mêmes lacunes que les précédentes par rapport à la spécificité du domaine et sont pénalisées par le nombre de requêtes nécessaires à effectuer pour calculer l'orientation sémantique de chaque mot (généralement au moins 29 requêtes).

3. L'Approche DOMA (Automatic Mining Opinion Dictionary)

L'objectif de cette section est de proposer un aperçu de l'approche DOMA. Le processus général est décrit dans la figure 1.

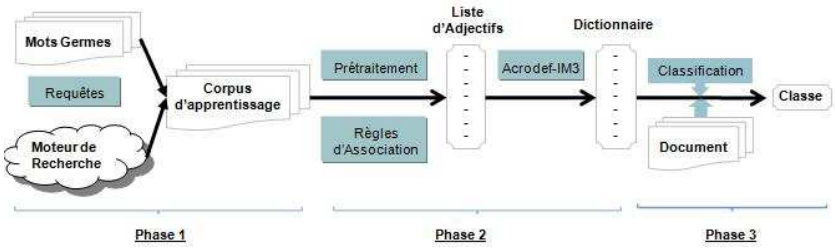

Figure 1. *Le processus général de l'approche* DOMA

Il est composé de trois phases :

– **Phase 1 : Acquisition du corpus d'apprentissage.** L'objectif de cette phase est d'extraire du Web de manière automatique des documents d'opinions exprimant des avis positifs ou négatifs.

– **Phase 2 : Extraction des adjectifs porteurs d'opinions.** Dans cette phase, nous recherchons les adjectifs positifs (resp. négatifs) associés à un ensemble d'adjectifs germes initiaux.

– **Phase 3 : Classification.** Cette phase a pour but de valider l'utilité des mots appris dans les deux phases précédentes utilisé pour la classification des documents.

Dans les sous sections suivantes, nous présentons en détail ces différentes phases.

3.1. *Phase 1 : Acquisition du Corpus d'Apprentissage*

Pour construire un dictionnaire d'opinion, la premiere étape consiste à acquérir de manière automatique un corpus adapté. Pour cela, nous considérons deux ensembles

P et N de mots "germes" dont les orientations sémantiques sont respectivement positif et négatif (Turney, 2002).

P = $\{good, nice, excellent, positive, fortunate, correct, superior\}$

N = $\{bad, nasty, poor, negative, unfortunate, wrong, inferior\}$

Pour chaque mot germe, nous utilisons un moteur de recherche avec une requête spécifiant un domaine d'application d, le mot germe recherché et les mots à éviter. Par exemple, si nous considérons le moteur de recherche google, pour obtenir des opinions sur des films avec le mot germe "good", la requête suivante est effectuée : "+opinion +review +cinema +good -bad -nasty -poor -negative -unfortunate -wrong -inferior". Cette requête donnera comme résultat des documents d'opinions sur le cinéma contenant le mot good mais ne contenant pas les mots bad, nasty, poor, ... inferior. De manière générale, les requêtes propres aux opinions positives (resp. négatives) d'un domaine d sont de la forme : "+opinion +mots$_d$ +germes$_{pos}$ -germes$_{neg}$. Dans le cadre de nos expérimentations (C.f. section 4), nous avons utilisé le moteur de recherche BlogGooglesearch.com spécialisé dans la recherche sur les blogs pour obtenir nos jeux d'apprentissage. Ainsi, pour chaque mot germe de l'ensemble P (resp. N) et pour un domaine donné, nous collectons automatiquement K documents où il n'apparaît aucun mot de l'ensemble N (resp. P). Nous obtenons ainsi, 14 corpus d'apprentissage partiels : 7 positifs et 7 négatifs.

3.2. *Phase 2 : Extraction des Adjectifs Porteurs d'Opinion*

Les corpus obtenus lors de l'étape précédente ne contiennent que des documents correspondant à un domaine spécifique. L'objectif de la seconde phase est de rechercher dans ces corpus les adjectifs porteurs d'opinion. Pour cela, à partir des corpus collectés, nous cherchons les corrélations entre les mots germes et les adjectifs des documents collectés pour enrichir les ensembles de mots germes avec des adjectifs pertinents. Cependant, pour éviter les faux positifs ou faux négatifs nous ajoutons une étape de filtrage. Nous présentons dans les sous sections suivantes ces étapes.

3.2.1. *Prétraitement et Règles d'Association*

Tout d'abord, nous utilisons « Tree Tagger » (Schmid, 1994). Comme dans (Taboada *et al.*, 2006, Voll *et al.*, 2007, Hatzivassiloglou *et al.*, 1997, Kamps *et al.*, 2004), nous considérons les adjectifs comme des mots représentatifs pour déterminer l'opinion. Ainsi, via TreeTagger nous ne conservons que les adjectifs des documents traités. Nous déterminons l'association sémantique entre les termes des documents et les mots germes des ensembles positifs et négatifs à l'aide d'un algorithme de recherche de règles d'association de type "Apriori" (Agrawal *et al.*, 1994). Soit $I = \{i_1, ...i_n\}$ un ensemble d'items, et D un ensemble de transactions, où chaque transaction correspond à un sous-ensemble d'éléments de I. Une règle d'association est une implication de la forme X$\rightarrow$Y, où X$\subset I$, Y$\subset I$, et X $\cap$ Y = $\oslash$. Une règle a un support s si s% des transactions de D contiennent X$\cup$Y. La règle X$\rightarrow$Y a une confiance c, si c% des tran-

sactions de D qui contiennent X contiennent aussi Y. Dans notre contexte, les items correspondent aux adjectifs et les transactions aux phrases.

3.2.2. *Filtrage*

De manière à minimiser le nombre de faux positifs et de faux négatifs, les adjectifs trouvés dans les documents qui sont en corrélation avec un seul mot germe sont supprimées. En effet, comme nous ne voulons pas utiliser de dictionnaire extérieur, en ne retenant que des adjectifs corrélés à plus d'un mot germe, nous souhaitons véritablement rechercher ceux qui sont fortement corrélés à des opinions positives ou négatives. Ensuite, pour les adjectifs qui apparaissent à la fois dans les listes positives et négatives, ceux qui sont corrélés avec plusieurs mots germes d'une même orientation ayant un support élevé et une moyenne d'apparition plus grande que 1, sont retenus comme mots appris de la part de ces mots germes, autrement ils sont éliminés.

Enfin, de manière à améliorer la qualité des résultats obtenus, nous appliquons la mesure AcroDef$_{MI3}$ décrite dans (Roche *et al.,* 2007). Cette approche consiste à mesurer la dépendance entre un adjectif de la liste des adjectifs appris par rapport à la liste des mots germes.

3.3. *Phase 3 : Classification*

Pour chaque document à classer, nous calculons son orientation positive ou négative en fonction du nombre d'adjectifs, appris dans la phase précédente, contenus dans le document. Nous comptons le nombre d'adjectifs positifs, puis le nombre d'adjectifs négatifs, et nous faisons la différence. Si le résultat est positif (resp. négatif), le document sera classé dans la classe positive (resp. négative). Autrement, il sera considéré comme neutre.

4. Expérimentations

Dans cette section, nous présentons les différentes expérimentations que nous avons réalisées pour valider notre méthodologie. Tout d'abord, nous présentons l'apprentissage des adjectifs puis les résultats obtenus lors de la classification.

Les documents que nous avons acquis via le moteur de recherche BlogGoogle-search.com sont relatifs aux opinions exprimées sur le cinéma. Les mots germes et les requêtes exécutées correspondent à ceux décrits dans la section 3.1. Pour chaque mot germe, nous avons limité le nombre de documents retournés par le moteur de recherche à 300. Nous transformons ensuite ces documents, du format HTML au format texte et utilisons TreeTagger pour ne retenir que les adjectifs.

De manière à étudier quelles sont les meilleures distances entre les mots germes et les adjectifs à retenir, nous avons effectué différentes expérimentations en faisant varier le paramètre Window Size (WS) de 1 à 3. WS=1 correspond au fait que nous retenons un adjectif avant le mot germe et un après. Ensuite, pour extraire les relations

de corrélation entre les adjectifs, nous utilisons une implémentation de l'algorithme Apriori[1]. Dans les expérimentations réalisées, nous avons fait varier le support de 1 à 3%. Nous obtenons ainsi pour chaque support, deux listes : une positive et une négative. Comme nous l'avons expliqué dans la section précédente, nous éliminons de ces listes les adjectifs en communs (pour une même valeur de support) ainsi que ceux qui sont en corrélation avec un seul mot germe. Pour supprimer les adjectifs fréquents inutiles, nous utilisons la mesure AcroDef$_{MI3}$ avec un seuil de 0.005 déterminé expérimentalement.

De manière à valider les connaissances acquises, nous utilisons le jeu de données Movie Review Data du NLP Group de l'Université de Cornell[2]. Ce jeu de données contient 1000 avis positifs et 1000 négatifs extraits de l'Internet Movie Database[3]. Lors des premières expérimentations, la classification est simplement réalisée en comparant le nombre d'adjectifs positifs et négatifs dans un texte. Le tableau 1 (resp. tableau 2) décrit la classification obtenue à partir des mots germes sans appliquer la méthode d'apprentissage sur le corpus de documents positifs (resp. négatifs). Dans le tableau 3 (resp. tableau 4), nous reportons les résultats obtenus à l'aide des adjectifs appris lors de la classification de documents positifs (resp. négatifs). La colonne WS correspond aux distances, la colonne S aux différentes valeurs de supports et, LP et LN correspondent aux nombres de vocabulaires positifs et négatifs. Par exemple, la valeur 7 + 12 de la colonne LP de la première ligne indique qu'il y a 7 adjectifs germes et 12 adjectifs appris. Comme nous pouvons le constater, notre méthode permet, dans le cas

Liste	Positifs	LP	LN
L.Germes	**66,9**%	7	7

Tableau 1. *Classification de 1000 documents positifs avec les mots germes*

Liste	Négatifs	LP	LN
L.Germes	**30,4**%	7	7

Tableau 2. *Classification de 1000 documents négatifs avec les mots germes*

WS	S	Positif	LP	LN
1	1%	**67,2**%	7+12	7+20
	2%	60,3%	7+8	7+13
	3%	65,6%	7+6	7+1
2	1%	57,6%	7+13	7+35
	2%	56,8%	7+8	7+17
	3%	68,4%	7+4	7+4
3	1%	28,9%	7+11	7+48
	2%	59,3%	7+4	7+22
	3%	67,3%	7+5	7+11

Tableau 3. *Classification de 1000 documents positifs avec les mots appris*

WS	S	Négatif	LP	LN
1	1%	**39,2**%	7+12	7+20
	2%	46,5%	7+8	7+13
	3%	17,7%	7+6	7+1
2	1%	49,2%	7+13	7+35
	2%	49,8%	7+8	7+17
	3%	32,3%	7+4	7+4
3	1%	76,0%	7+11	7+48
	2%	46,7%	7+4	7+22
	3%	40,1%	7+5	7+11

Tableau 4. *Classification de 1000 documents négatifs avec les mots appris*

1. http ://fimi.cs.helsinki.fi/fimi03/
2. http ://www.cs.cornell.edu/people/pabo/movie-review-data/
3. http ://www.imdb.com/

des documents négatifs, une classification nettement améliorée. Dans le cas des documents positifs, la différence est cependant moins importante mais comme l'illustre le tableau 5, nous pouvons constater que les mots appris apparaissent de manière très significatives dans les documents de test.

Comme cela était prévisible les meilleurs résultats, en comparant le nombre d'adjectifs appris, ont été obtenus pour un WS de 1. En effet, le WS de 1 indique que nous nous intéressons plus particulièrement aux adjectifs localisés juste avant et juste après le mot germe. Cette expérience confirme les hypothèses de proximité d'adjectifs dans l'expression d'opinions (Turney, 2002), nous pouvons le constater dans les tableaux 3 et 4. Le nombre d'adjectifs positifs et négatifs appris en fonction des valeurs de support peut varier très fortement. Par exemple, pour un support de 1% et WS=3, nous voyons qu'il y a 11 adjectifs positifs appris et 48 négatifs. Une analyse fine des résultats a effectivement montré que la plupart de ces derniers correspondaient à des adjectifs fréquents inutiles. Les résultats obtenus en appliquant la mesure AcroDef$_{MI3}$ qui consiste à filtrer les adjectifs sont décrits dans les tableaux 6 et 7, où nous ne reportons que les valeurs pour WS=1 et S=1%. Nous constatons que le pourcentage de documents bien classés, grâce à notre approche, passe pour les positifs de 66.9% à 75.9% et pour les négatifs de 30.4% à 57.1%.

Germes positifs	
Adjectif	Nb d'occ.
Good	2147
Nice	184
Excellent	146
Superior	37
Positive	29
Correct	27
Fortunate	7

Adjectifs positifs appris			
Adjectif	Nb d'occ.	Adjectif	Nb d'occ.
Great	882	Hilarious	146
Funny	441	Happy	130
Perfect	244	Important	130
Beautiful	197	Amazing	117
Worth	164	Complete	101
Major	163	Helpful	52

Tableau 5. *Occurrences des adjectifs positifs pour WS=1 et S=1%*

WS	S	Positif	LP	LN
1	1%	**75,9**%	7+11	7+11

Tableau 6. *Classification de 1000 documents positifs avec mots appris et application d'AcroDef$_{MI3}$*

WS	S	Négatif	LP	LN
1	1%	**57,1**%	7+11	7+11

Tableau 7. *Classification de 1000 documents négatifs avec les mots appris et application d'AcroDef$_{MI3}$*

5. Conclusion

Dans cet article, nous avons proposé une nouvelle approche de détection automatique d'adjectifs positifs et négatifs pour la fouille de données d'opinions. Les résultats menés sur des jeux de données issus de l'apprentissage (blogs vs. reviews de cinéma)

ont montré que via notre approche nous étions capables d'apprendre les adjectifs pertinents.

Les perspectives à ce travail sont nombreuses. Tout d'abord même si nous ne l'avons pas évoqué dans cet article, l'approche d'apprentissage proposée peut bien entendu être étendue en renforçant l'apprentissage. Dans ce cas, il suffit simplement de reconsidérer les mots acquis comme de nouveaux mots germes. La seconde perspective est liée à la technique de classification qui pour l'instant est assez sommaire : présence ou absence de mots. Nous souhaitons donc améliorer cette étape en intégrant les travaux récents issus de la linguistique, comme la prise en compte des adverbes inversant la polarité. Enfin, nous souhaitons appliquer l'approche à d'autres domaines afin de vérifier que les choix effectués dans la spécification des paramètres restent valides et donc généralisables.

6. Bibliographie

Agrawal R., Srikant R., « Fast Algorithms for Mining Association Rules in Large Databases », *VLDB'94*, 1994.

Andreevskaia A., Bergler S., « Semantic Tag Extraction from WordNet Glosses », 2007.

Hatzivassiloglou V., McKeown K., « Predicting the semantic orientation of adjectives », *In Proceedings of 35th Meeting of the Association for Computational Linguistics*, 1997.

Hu M., Liu B., « Mining and Summarizing Customer Reviews », *In Proceedings of KDD'04, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

Kamps J., Marx M., Mokken R. J., de Rijke M., « Using WordNet to Measure Semantic Orientation of Adjectives », *In Proceedings of LREC 2004, the 4th International Conference on Language Resources and Evaluation*, vol. IV, p. 174-181, 2004.

Miller G., « WordNet : A Lexical database for English », *Communications of the ACM*, 1995.

Plantié M., Roche M., Dray G., « Un système de vote pour la classification de textes d'opinion », *Proceedings of the 8th journées francophones d'Extraction et gestion des Connaissances*, 2008.

Roche M., Prince V., « Acrodef : A Quality Measure for Discriminating Expansions of Ambiguous Acronyms », *CONTEXT*p. 411-427, 2007.

Schmid H., « TreeTagger », *TC project at the Institute for Computational Linguistics of the University of Stuttgart*, 1994.

Stone P., Dunphy D., Smith M., Ogilvie D., « The General Inquirer : A Computer Approach to Content Analysis », 1966.

Taboada M., Anthony C., , Voll K., « Creating semantic orientation dictionaries », 2006.

Turney P., « Thumbs up or thumbs down ? Semantic orientation applied to unsupervised classification of reviews », *In Proceedings of 40th Meeting of the Association for Computational Linguistics*p. 417-424, 2002.

Turney P., Littman M., « Measuring praise and criticism : Inference of semantic orientation from association », *ACM Transactions on Information Systems*p. 315-346, 2003.

Voll K., Taboada M., « Not All Words are Created Equal : Extracting Semantic Orientation as a Function of Adjective Relevance », 2007.