Détection d'opinion. Comment déterminer les adjectifs d'opinion d'un domaine donné

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

▶ **To cite this version:**

Ali Harb, Michel Plantié, Mathieu Roche, Gérard Dray, François Trousset, et al.. Détection d'opinion. Comment déterminer les adjectifs d'opinion d'un domaine donné. Document numérique - Revue des sciences et technologies de l'information. Série Document numérique, 2008, 11 (1-2), pp.37-61. 10.3166/DN.11.1-2.37-61 . lirmm-00349236

HAL Id: lirmm-00349236

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00349236v1

Submitted on 5 Apr 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détection d'opinion

Comment déterminer les adjectifs d'opinion d'un domaine donné

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

** EMA-LGI2P, Parc Scientifique Georges Besse*
F-30035 Nîmes cedex
{ali.harb, michel.plantie, gerard.dray, francois.trousset, pascal.poncelet}@ema.fr

*** LIRMM Université Montpellier II*
CNRS 5506, 161 Rue Ada
F-34392 Montpellier
{ali.harb, mathieu.roche}@lirmm.fr

RÉSUMÉ. *L'extraction automatique d'opinions sur le web 2.0 est un domaine de recherche de plus en plus étudié. Elle utilise souvent deux méthodes à vocations différentes : soit des méthodes fondées sur l'apprentissage par la constitution de corpus en vue d'établir des modèles pour la classification, soit rechercher des mots caractéristiques tels que les adjectifs qui contribueront à la classification des textes. Dans ce dernier cas, les outils existants utilisent des dictionnaires généraux, et possèdent des 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 intégrant la connaissance du domaine. Les expériences menées sur des données réelles ont montré l'intérêt de notre approche comparativement à une méthode plus classique par apprentissage.*

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, compared to a more classic method based on machine learning mechanisms.*

MOTS-CLÉS : *fouille de texte, règles d'association, orientation sémantique, classification.*

KEYWORDS: *text mining, association rules, semantic orientation, classification.*

DOI:10.3166/DN.11.1-2.37-61 © 2008 Lavoisier, Paris

DN – 11/2008. Documents et web 2.0, pages 37 à 61

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

1. Introduction

Avec le développement du web, et surtout du web 2.0, le nombre de documents décrivant des opinions sur un produit ou un film devient de plus en plus important. Récemment, les chercheurs de différentes communautés (Fouille de données, Fouille de textes, Linguistique) se sont intéressés à l'extraction automatique de ces données d'opinions sur le web. Certaines techniques de détection d'opinions cherchent à déterminer les caractéristiques d'opinions positives ou négatives à partir d'ensembles d'apprentissages. Des experts sont mandatés pour constituer des corpus de référence, et des techniques de classification (se fondant notamment sur différentes techniques linguistiques) sont alors utilisées pour classer automatiquement les documents extraits du web. Dans cet article, nous nous intéressons aux techniques fondées sur l'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, la plupart des approches utilisent des dictionnaires existants ou des listes prédéfinies d'adjectifs. Dans ce cas, elles 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 pour détecter automatiquement les adjectifs correspondant à une opinion exprimée dans un domaine spécifique.

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

2. Travaux antérieurs

2.1. *Méthodes supervisées fondées sur l'existence de corpus d'opinion*

Les méthodes supervisées reposent sur l'existence préalable de corpus d'opinion constitués par des experts du domaine. L'avantage est alors de pouvoir utiliser des techniques de fouille de textes combinant outils linguistiques et outils de classification pour déterminer l'opinion d'un nouveau texte. L'idée est d'apprendre automatiquement les unités linguistiques ou termes au sens large pour modéliser une opinion particulière. Les termes extraits sont dépendants du domaine considéré. Puis des techniques à base de différentes méthodes de classification sont utilisées (Plantié *et al.*, 2008; Plantié, 2006). Ces méthodes sont souvent utilisées dans des challenges nationaux (Grouin *et al.*, 2007) et internationaux (Yang *et al.*, 2006). Si l'on dispose de corpus d'apprentissage bien structurés alors ces méthodes d'apprentissage supervi-

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

sées donnent d'excellents résultats. Cependant la difficulté réside dans la constitution de ces corpus d'apprentissage, qui est un processus manuel à effectuer pour chaque domaine étudié.

2.2. *Méthodes non supervisées de détection d'opinion*

Comme nous l'avons mentionné précédemment, la plupart des approches non supervisées 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 (par exemple Inquirer (Stone *et al.*, 1966) contient 3 596 mots étiquetés positifs ou négatifs ou HM (Hatzivassiloglou *et al.*, 1997) répertorie 1 336 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). Avec ces méthodes, la qualité du résultat final est fortement liée aux différents dictionnaires disponibles, malheureusement ces méthodes 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 statistiques 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.

3. L'Approche AMOD (*Automatic Mining Opinion Dictionary*)

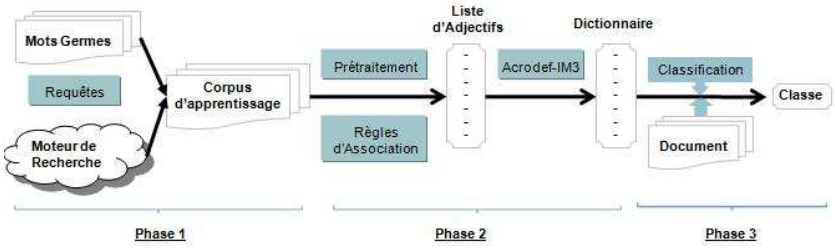

Figure 1. *Le processus général de l'approche* AMOD

L'objectif de cette section est de présenter l'approche AMOD. Le processus général est décrit dans la figure 1.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Il est composé de trois phases :

– **Phase 1 : acquisition du corpus d'apprentissage.** L'objectif de cette phase est d'extraire de manière automatique du web 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 à partir du corpus d'apprentissage,

– **Phase 3 : classification.** Cette phase a pour but de valider l'utilité des adjectifs appris dans les deux phases précédentes en classifiant de manière automatique des documents.

Dans les 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 première étape consiste à acquérir un corpus adapté, de manière automatique. Pour cela, nous considérons deux ensembles P et N de mots *germes* classiquement utilisés dans la littérature dont les orientations sémantiques sont respectivement positives et négatives (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 des mots à éviter. Par exemple, si nous considérons le moteur de recherche Google, pour obtenir des corpus d'opinions sur des films avec le mot germe « good », la requête suivante est effectuée : *« +opinion +review +movies +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... inferior*.

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, après avoir converti les documents du format « HTML » au format « TEXT », 14 corpus de documents correspondant chacun à un mot germe : 7 positifs et 7 négatifs. Les 7 ensembles de documents associés aux mots germes positifs (resp. négatifs) représentent le corpus d'apprentissage positif (resp. négatif).

Nous pouvons représenter cette phase d'acquisition de corpus avec l'algorithme 1. Pour chaque mot germe *p* de l'ensemble *P*, nous générons une requête *R* composée d'un moteur de recherche *M*, d'un domaine (*i.e.* contexte) *d*, d'un ensemble de mots germes N à éliminer. A partir de cette requête, nous collectons automatiquement *K* documents (*fonction get(R,K)*). Pour chaque document, nous appliquons la fonction

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

convert() qui convertit du format *HTML* au format *TEXT*. Ces *K* documents convertis construisent un corpus relatif aux mots germes « *p* ». Nous procédons de la même manière pour les mots germes négatifs.

Algorithm 1: *Création de corpus d'apprentissage*

Input: Le moteur de recherche M, le domaine d'intérêt d, les ensembles de mots germes positifs et négatifs P et N

Output: Les corpus d'apprentissage Positifs et Négatifs C_P

```
begin
    C_P = ∅
    foreach p in P do
        R = : <M, d, p, N> ;
        C_p = get(R, K) ;
        foreach D_i in C_p do
            Convert(D_i) ;
        C_P = C_P ∪ C_p ;
end
```

3.2. *Phase 2 : extraction des adjectifs porteurs d'opinions*

Les corpus obtenus lors de l'étape précédente contiennent des documents correspondant à un domaine spécifique et porteurs d'opinions. L'objectif de la seconde phase est de rechercher dans ces corpus les adjectifs spécifiques au domaine et porteurs d'opinions. Pour cela, à partir des corpus collectés, nous cherchons des corrélations entre les mots germes et d'autres adjectifs dans les documents collectés. Le but est d'enrichir les ensembles de mots germes par des adjectifs pertinents et utiles. Cependant, comme nous le verrons ci-après, ce processus collecte également des adjectifs peu discriminants. Pour éviter ce défaut, nous ajoutons une étape de filtrage. Nous présentons dans les sous-sections ci-dessous ces deux étapes (extraction des règles d'association et filtrage).

3.2.1. *Prétraitement et règles d'association*

Afin d'établir des associations entre différents adjectifs pour enrichir un dictionnaire d'opinion, il est tout d'abord nécessaire de connaître la fonction grammaticale de chacun des mots de notre corpus d'apprentissage. Pour ce faire, nous utilisons l'outil Tree Tagger (Schmid, 1994). Ce système d'étiquetage automatique de textes attribue à chaque mot une catégorie grammaticale et fournit les mots sous une forme lemmatisée (forme canonique). Les règles d'étiquetage du Tree Tagger sont apprises en appliquant un algorithme d'arbre de décision (Quinlan, 1986) à partir d'un corpus d'apprentissage étiqueté manuellement. Nous montrons ci-après, un exemple de sortie du Tree Tagger à partir du texte suivant :

On ne change pas une équipe qui gagne.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

La figure 2 correspondant au résultat donné par l'outil Tree Tagger montre trois types d'informations. Le premier est le mot lui même, tel qu'il est trouvé dans le texte original. Ensuite la fonction grammaticale des mots est donnée (e.g. *PRO : PER, décrit le pronom personnel, ADV adverbe, VER : PRES verbe conjugué au présent, DET : ART désigne un article, SENT ponctuation qui désigne la fin d'une phrase*). La dernière colonne correspond au lemme associé au mot d'origine.

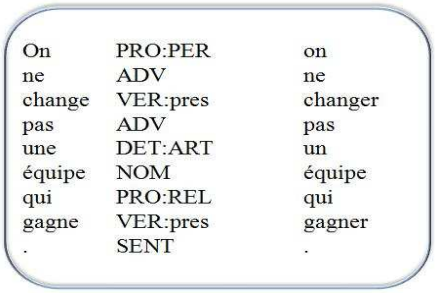

Figure 2. *Exemple du fichier généré par Tree Tagger*

Ainsi nous utiliserons l'outil *Tree Tagger* sur nos corpus d'apprentissage dans le but d'en extraire les mots particulièrement porteurs d'opinion tels que les adjectifs (Taboada *et al.*, 2006; Voll *et al.*, 2007; Hatzivassiloglou *et al.*, 1997; Strapparava *et al.*, 2004; Esuli *et al.*, 2005). L'étape suivante consiste alors à déterminer l'association entre les termes (ici les adjectifs) des documents et les mots germes des ensembles positifs et négatifs. Le but est de déterminer si les adjectifs trouvés sont porteurs des mêmes opinions que les mots germes. Pour cela nous utilisons un processus d'extraction de règles d'association.

En effet, le principe général des règles d'association est de rechercher des corrélations entre des items stockés dans une base de données. Dans notre cas, il s'agit plus particulièrement de rechercher comment les adjectifs sont corrélés entre eux.

Nous rappelons les principes de l'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→Y, où X⊂I, Y⊂I, et X∩I $= \oslash$. Une règle a un support s si s % des transactions de D contiennent X∪Y. La règle X→Y a une confiance c, si c % des transactions de D qui supportent X supportent Y.

Dans notre contexte, les items correspondent aux adjectifs et les transactions aux phrases. Les transactions sont créées à partir des fenêtres de type (*WS*) composées par des adjectifs où les mots germes sont les pivots. Notons que les adjectifs sont identifiés à l'aide de l'étiqueteur « Tree Tagger ».

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Etant donné que dans le cadre des règles d'association, chaque transaction correspond à un ensemble d'items, notre objectif est de repérer au sein des phrases comment représenter ces transactions. Le principe retenu est le suivant : nous considérons *via* des fenêtres comment constituer des transactions. En faisant varier la taille des fenêtres nous souhaitons étudier à partir de quel moment les adjectifs restent pertinents. Pour cela, nous souhaitons faire varier WS de 1 à 3. Dans ce contexte, un WS=1 correspond à un adjectif avant et un adjectif après le mot germe.

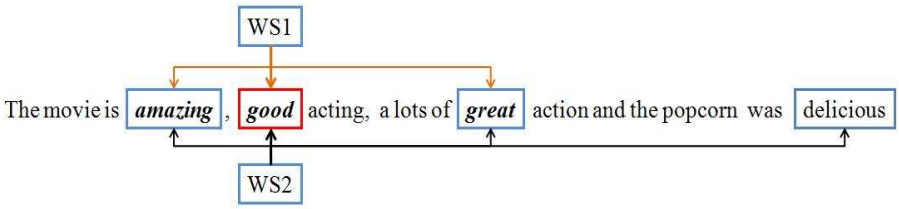

Figure 3. *Exemple de « Window Size » (WS)*

Dans la figure 3, le pivot est *good*. En utilisant *WS* de taille 1, nous obtenons les adjectifs *amazing* et *great*. Par contre en spécifiant un *WS* de taille 2, l'adjectif *delicious* est ajouté à la transaction. Plus la taille des fenêtres *WS* est grande, plus l'association des adjectifs collectés s'éloigne de celle du mot germe.

3.2.2. *Filtrage*

Etant donné que notre objectif est de rechercher de manière automatique des adjectifs qui soient dépendants des domaines d'applications, nous souhaitons que ceux-ci soient les plus représentatifs possible. Bien entendu, *via* l'étape précédente, nous obtenons l'ensemble des adjectifs (dans une fenêtre donnée) qui sont fortement corrélés. Toutefois, de manière à extraire ceux qui sont en forte corrélation, nous supprimons du résultat précédent, ceux qui ne sont corrélés qu'avec un seul mot germe. La deuxième étape de filtrage examine les adjectifs apparaissant à la fois dans les listes positives et négatives. Ceux d'entre eux qui sont corrélés avec plusieurs mots germes d'une même orientation et ayant un support élevé sont analysés. Si leur fréquence d'apparition dans un corpus d'une des deux orientations est supérieure à 1 occurrence par document, ils sont retenus comme mots appris à partir de ces mots germes. Dans le cas contraire ils sont éliminés. Remarque : après les traitements de règles d'association, il est impossible de trouver des adjectifs ayant une fréquence d'apparition supérieure à 1 par document, à la fois dans les corpus positifs et négatifs.

A l'issue de cette phase, nous obtenons donc des listes d'adjectifs positifs et négatifs qui peuvent se révéler bruités. Ceci est dû aux données issues du corpus qui peuvent parfois se révéler non adaptées.

Ainsi, de manière à améliorer la qualité des résultats obtenus, nous appliquons la mesure $AcroDef_{IM3}$ décrite dans (Roche *et al.*, 2007). Le but est alors de mesurer la

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

« force » d'association entre les mots germes et les mots associés trouvés par les règles d'association. Pour ce faire nous proposons de mesurer la dépendance de deux mots (mots germe et mot appris) dans un contexte donné.

AcroDef$_{IM3}$ s'appuie sur des données statistiques fournies par le résultat de moteurs de recherche. Cette mesure est fondée sur l'Information Mutuelle au cube (Daille, 1994) en intégrant une notion de contexte. L'Information Mutuelle au cube est une mesure empirique fondée sur l'Information Mutuelle (Church *et al.*, 1989) qui calcule une certaine forme de dépendance entre les mots x et y :

$$IM3\,(x,y) = log_2\left(\frac{nb(x,y)^3}{nb(x)\times nb(y)}\right)$$

Contrairement à l'information mutuelle, l'information mutuelle au cube privilégie les co-occurrences fréquentes. Par ailleurs, *AcroDef*$_{IM3}$ prend en considération le contexte dans lequel les co-occurrences sont présentes. En appliquant un contexte C (représenté par un ensemble de mots), l'approche *AcroDef*$_{IM3}$ est donnée par la formule suivante :

$$\textit{AcroDef}_{IM3}(x,y) = log_2\left(\frac{nb((x,y)\ and\ C)^3}{nb(x\ and\ C)\times nb(y\ and\ C)}\right)$$

Dans nos travaux, nous souhaitons calculer la dépendance entre deux adjectifs x, y. Dans ce cas, nb(x,y) représente le nombre de pages web où x et y sont présents ensemble et de manière consécutive. Ainsi, dans la pratique, nb(x,y) correspond au nombre total de pages retournées par un moteur de recherche avec les deux requêtes « x y » et « y x » (somme totale du nombre de pages retournées par ces deux requêtes). Ceci permet de déterminer les pages citant deux adjectifs voisins (par exemple, dans le cas d'énumérations). En considérant, un contexte *C* représenté par les mots $\{c_1, ..., c_n\}$, les requêtes associées sont :

$$\{\text{"}xy\text{"}\ AND\ c_1\ AND ...\ AND\ c_n\}\ et\ \{\text{"}yx\text{"}\ AND\ c_1\ AND\ \ ...\ AND\ c_n\}$$

Considérons l'adjectif extrait *funny* et classé comme ayant une orientation sémantique positive. Le mot-clé de notre contexte est *movies*. Nous allons chercher la dépendance entre l'adjectif *funny* et les mots germes positifs cités dans la section 3.1 pour le contexte *movies*. Prenons l'adjectif *good*. Par rapport à l'adjectif *funny* et le mot germe *good* dans le contexte cinéma, nous obtenons la formule *AcroDef*$_{IM3}$ suivante :

$$\textit{AcroDef}_{IM3}(funny, good) =$$

$$log_2\left(\frac{(nb((\text{'}\ funny,good\ \text{`})\ and\ \text{'}\ movies\ \text{`})+nb((\text{'}\ good,funny\ \text{`})\ and\ \text{'}\ movies\ \text{`}))^3}{nb(funny\ and\ movies)\times nb(good\ and\ movies)}\right)$$

Cette formule est appliquée en utilisant le moteur de recherche Google, et nous obtenons la valeur *118.48*. La même formule est appliquée entre l'adjectif *funny* et

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

tous les mots germes (voir figure 4). Ensuite nous calculons la moyenne de toutes les valeurs obtenues (*17.37* dans l'exemple de la figure 4). Puisque, cette valeur est supérieure au seuil de 0,005 déterminé expérimentalement, l'adjectif *funny* sera retenu comme adjectif positif appris.

3.2.3. *Synthèse : algorithme de construction d'un dictionnaire d'opinion spécifique à partir d'un corpus*

Cette section résume l'algorithme de construction de dictionnaire d'opinion à partir d'un corpus, et présente le déroulement d'un tel algorithme avec un exemple issu du domaine du *cinéma*.

Algorithm 2: *Création des Fenêtres(WS) et Génération des Régles d'Association*

Input: Corpus d'apprentissage C_P, support minimum sup, taille de la fenêtre (Window Size) w, seuil d'*Acrodef*$_{IM3}$ s

Output: Ensemble Positif de règles d'association L_P

begin
- **foreach** C_p *in* C_P **do**
  - **foreach** D_i *in* C_p **do**
    - D_T = *TreeTagger*(D_i)
    - $C'_p = C'_p \cup$ *Window*(D_T, w) ;
  - l_p = *Apriori*(C'_p, sup) ;
  - $L_P = L_P \cup l_p$;
- *Filter_com*(L_P) ;
- *Acrodef*$_{IM3}$(L_P, s)

end

Pour chaque corpus partiel C_p, nous appliquons *TreeTagger* pour lemmatiser chaque document, et obtenir les fonctions grammaticales des mots. Ensuite nous appliquons la fonction *Window* qui créé les transactions en fonction des mots germes et des adjectifs. Nous obtenons un nouveau corpus C'_p. Sur ce nouveau corpus, nous appliquons l'algorithme de règles d'association « Apriori » vu précédemment avec un support minimum sup pour générer les règles d'association. Nous obtenons la liste partielle L_p de règles d'association entre les adjectifs de chaque corpus partiel et les mots germes. Les listes de règles d'associations de tous les corpus partiels positifs sont fusionnées pour former la liste globale des règles d'associations L_P. La même procédure est appliquée sur les corpus partiels négatifs pour obtenir la liste globale des règles d'associations L_N. Nous donnons un exemple de règle dans la figure 5.

De manière à illustrer l'étape d'extraction des adjectifs porteurs d'opinions, considérons les règles d'associations présentées dans la figure 5. Ces règles sont extraites en appliquant l'algorithme « Apriori » sur le résultat des transactions effectuées pour constituer le corpus d'apprentissage. Tout d'abord, le processus détecte et élimine les adjectifs communs aux deux corpus et n'ayant pas une fréquence d'apparition suffisante dans au moins un des deux corpus (par exemple pour l'adjectif « différent »).

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

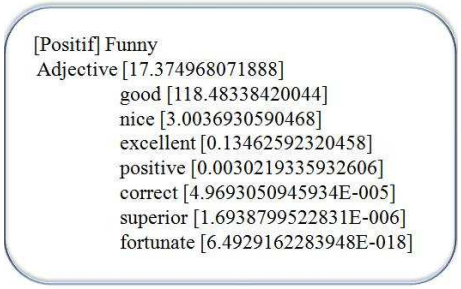

Figure 4. *Exemple funny en appliquant AcroDef$_{IM3}$*

Règles d'associations	
Positifs	Négatifs
1. good → funny	1. bad→ boring
2. good → great	2. bad → commercial
3. good → ***different***	3. bad → ***different***
4. excellent → funny	4. wrong → boring
5. nice → great	5. wrong → commercial
6. nice → encouraging	6. poor → current

Figure 5. *Exemple de règles d'associations*

Nous obtenons deux listes d'adjectifs positifs et négatifs. La figure 6 montre les adjectifs candidats, et à côté de chacun d'eux, nous trouvons les adjectifs germe avec lesquels ils sont en corrélation (e.g. funny est en corrélation avec good et excellent). Pour éliminer les adjectifs non pertinents, nous appliquons la mesure de dépendance *AcroDef*$_{IM3}$ avec un seuil déterminé expérimentalement.

Filtrages de commun	
Positifs	Négatifs
1. funny : good, excellent	1. boring : bad, wrong
2. great : good, nice	2. commercial : bad, wrong
3. encouraging : nice	3. current : poor

Figure 6. *Listes des adjectifs positifs après suppression des adjectifs communs*

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Dans la figure 7, nous remarquons que les adjectifs *encouraging* et *current* ont une valeur de dépendance 0,001 et 0,002. Si ces valeurs sont inférieures au seuil de *0,005*, les adjectifs sont supprimés.

Filtrages avec $AcroDef_{IM3}$	
Positifs	Négatifs
1. *funny* (20,948)	1. *boring* (8,330)
2. *great* (12,529)	2. *commercial* (3,054)
3. ~~*encouraging*~~ (0,001)	3. ~~*current*~~ (0,0002)

Figure 7. *Liste des adjectifs retenus après l'application de AcroDef$_{IM3}$*

3.3. *Phase 3 : classification*

Même si cette section ne représente pas le cœur de notre contribution, nous en faisons une description rapide. Pour chaque document à classer, nous calculons son orientation positive ou négative, en fonction du nombre d'adjectifs appris dans la phase précédente contenu dans le document. Si le résultat est positif (resp. négatif) le document sera classé dans la classe positive (resp. négative). Nous avons cependant étendu cette méthode pour prendre en compte les adverbes qui inversent la polarité. Pour améliorer la classification, nous analysons morpho-syntaxiquement nos documents pour détecter les négations. Nous avons cherché les adverbes associés directement à des adjectifs (par exemple : not, neither, nor), et nous avons alors modifié le degré d'orientation sémantique de ces adjectifs.

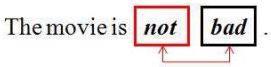

Figure 8. *Exemple d'utilisation de la négation*

Prenons l'exemple de la figure 8, si nous essayons de déduire l'orientation sémantique de l'adjectif *bad* en appliquant la méthode originale, nous remarquons qu'il a une orientation sémantique négative, et il influe sur l'orientation sémantique de la phrase en lui donnant une orientation négative. Par contre, si nous prenons en compte les adverbes qui influent sur la polarité des adjectifs, nous pouvons renverser la polarité de l'adjectif et éviter l'erreur d'affectation de polarité. Donc, en détectant *not* avant l'adjectif *bad*, l'effet donné par *not* inverse la polarité de *bad* qui devient positive.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

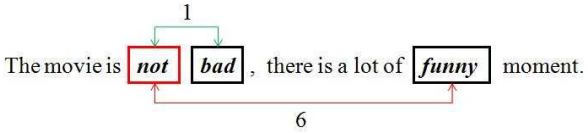

Figure 9. *Exemple d'utilisation de la forme négative*

Dans la figure 9, l'adverbe de négation *not* est adjacent à l'adjectif *bad* (*i.e.* distance égale à un) alors son orientation sémantique est inversée. Cependant nous considérons que l'inversion de polarité n'affecte que les adjectifs adjacents à un adverbe de négation. Pour cette raison-là, l'adjectif *funny* situé a une distance de *6* mots, ne verra pas sa polarité inversée par l'effet du *not* qui est ignoré. Il conserve son orientation sémantique initiale.

En ce qui concerne *not* et *neither nor*, nous avons traité sept cas :

1) Not ADJ
2) Not ADJ at all
3) Not very ADJ
4) Not so ADJ
5) Not too ADJ
6) Not ADJ enough
7) Neither ADJ_1 nor ADJ_2

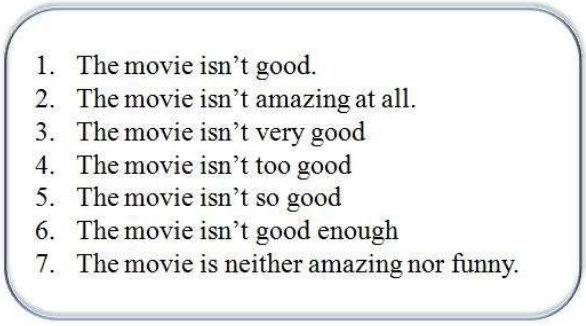

Figure 10. *Exemple de négation*

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

La procédure de calcul concernant les inversions ou modifications de polarités est indiquée ci-après. Nous calculons pour chaque adjectif *p*, appartenant à la liste des adjectifs positifs et qui apparaît dans le document, son nombre d'occurrences tout en respectant l'opérateur de négation. Dans un premier temps, pour chaque phrase *l* du document, nous testons si l'adjectif *p* est présent. Si c'est le cas, nous cherchons ensuite à détecter s'il y a un des adverbes de négation dans la phrase tout en respectant la distance entre ces adverbes et l'adjectif *p* (e.g. distance de 1). Si un tel adverbe existe, l'orientation sémantique de l'adjectif est inversée. Par contre s'il existe un adverbe comme *so*, *too* ou *very* entre le *not* et l'adjectif *p*, l'effet de cet adverbe se traduit en augmentant l'orientation sémantique de l'adjectif de 30 %. Par contre si nous trouvons *enough*, son orientation sémantique est diminuée de 30 %. L'orientation sémantique de chaque adjectif est initialisée à *1* dans nos expérimentations. De manière à illustrer ces cas, nous proposons pour chacun d'entre eux un exemple dans la figure 10. Pour les exemples 1, 2 et 7, nous remarquons que l'orientation sémantique des adjectifs *good* et *amazing* est totalement inversée. Par contre, elle est augmentée de 30 % de sa valeur initiale, pour l'adjectif *good* dans les exemples 3, 4 et 5, et est diminuée de 30 % pour le sixième exemple tout en respectant la distance entre l'adverbe et l'adjectif.

La somme des occurrences de tous les adjectifs de la liste positive représente l'orientation sémantique positive de ce document. Nous appliquons de la même manière cette procédure sur la liste négative. La différence de l'orientation sémantique positive et négative nous donne ainsi l'orientation sémantique du document.

4. Expérimentations

Dans cette section, nous présentons les différentes expérimentations que nous avons réalisées pour valider notre méthodologie. La constitution du corpus d'apprentissage des adjectifs est réalisée *via* le moteur de recherche BlogGooglesearch.com en considérant les opinions exprimées sur le domaine du cinéma. Les mots germes et les requêtes exécutées correspondent à ceux décrits dans la section 3.1. Pour chaque mot germe, le nombre de documents retourné par le moteur de recherche est de *300* (Cf. section 4.2). Tous les documents sont transformés en texte et l'utilisation de *Tree Tagger* permet de ne récupérer que les adjectifs. Pour le corpus de test nous utilisons le jeu de données Movie Review Data du NLP Group de l'Université de Cornell[1]. Ce jeu de données contient 1 000 avis positifs et 1 000 avis négatifs extraits de l'Internet Movie Database[2]. Les bases de données utilisées pour l'apprentissage et pour le test sont deux bases très différentes. La première correspond à des blogs alors que la seconde est issue de documents journalistiques. Ce choix a été fait afin d'étudier la robustesse de notre approche quand des opinions sont exprimées de manière différente. Nous reviendrons sur cet aspect lors de la comparaison avec une méthode de classification traditionnelle (Cf. section 4.3).

1. http ://www.cs.cornell.edu/people/pabo/movie-review-data/
2. http ://www.imdb.com/

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

La partie création de corpus d'apprentissage a été écrite en Perl et exécutée sur une machine Sun sous Unix. Les deux parties d'extraction et de classification ont été développées en Java et PHP, et exécutées sur un PC sous Windows XP. La partie règle d'association utilise une implémentation de l'algorithme Apriori[3].

4.1. *Evaluation*

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

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

Tableau 1. *Classification de 1 000 documents positifs et négatifs avec les mots germes*

Lors des premières expérimentations, la classification est simplement réalisée en comparant le nombre d'adjectifs positifs et négatifs dans un texte. Le tableau 1 décrit la classification obtenue à partir des mots germes sans appliquer notre méthode d'apprentissage sur le corpus de documents positifs et négatifs.

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

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 2. *Classification de 1 000 documents positifs et négatifs avec les mots appris*

Dans le tableau 2, nous reportons les résultats obtenus à l'aide des adjectifs appris lors de la classification de documents positifs (resp. négatifs). 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. La colonne WS correspond à ce paramètre et WS=1 correspond au fait que nous retenons un adjectif avant le mot germe et un après. Nous avons fait varier le support de 1 à 3 %. La colonne S correspond aux différentes valeurs de supports. LP et LN correspondent aux nombres d'adjectifs positifs et négatifs. Par exemple, la valeur 7 + 12 de la colonne LP (première ligne) indique qu'il y a 7 adjectifs germes et 12 adjectifs appris.

3. http ://fimi.cs.helsinki.fi/fimi03/

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Comme nous pouvons le constater, notre méthode permet, dans le cas des documents négatifs, une classification nettement améliorée. Dans le cas des documents positifs, la différence est cependant moins importante mais comme l'illustrent les tableaux 3 et 4, nous pouvons constater que les mots appris positifs et négatifs apparaissent de manière très significative 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. Cette expérience confirme les hypothèses de proximité d'adjectifs dans l'expression d'opinions proposée dans (Turney, 2002).

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

Adjectifs positifs appris			
Adjectifs	Nb d'occ.	Adjectifs	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 3. *Occurrences des adjectifs positifs pour WS=1 et S=1 %*

Germes négatifs	
Adjectifs	Nb d'occ.
Bad	1413
Wrong	212
Poor	152
Nasty	38
Unfortunate	25
Negative	22
Inferior	10

Adjectifs négatifs appris			
Adjectifs	Nb d'occ.	Adjectifs	Nb d'occ.
Boring	200	Certain	88
Different	146	Dirty	33
Ridiculous	117	Social	33
Dull	113	Favorite	29
Silly	97	Huge	27
Expensive	95		

Tableau 4. *Occurrences des adjectifs négatifs pour WS=1 et S=1 %*

Comme nous pouvons le constater dans le tableau 2, 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$_{IM3}$, avec un seuil de 0,005, pour filtrer les adjectifs sont décrits dans le tableau 5. 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 %**. Le tableau 6 illustre les adjectifs positifs et négatifs supprimés.

Dans les expérimentations suivantes, nous réappliquons le principe d'AMOD sur les adjectifs obtenus en considérant ces derniers comme des adjectifs germes afin de

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Résultat de 1 000 documents Positifs				
WS	S	Positif	LP	LN
1	1 %	**75,9** %	7+11	7+11
	2 %	46,2 %	7+6	7+8
	3 %	68,8 %	7+5	7+1
2	1 %	50,6 %	7+11	7+18
	2 %	44,1 %	7+6	7+11
	3 %	50,0 %	7+3	7+4
3	1 %	31,9 %	7+11	7+32
	2 %	48,5 %	7+4	7+15
	3 %	54,8 %	7+5	7+6
4	1 %	46,8 %	7+16	7+30
	2 %	35,7 %	7+6	7+25
	3 %	49,9 %	7+3	7+9

Résultat de 1 000 documents Négatifs				
WS	S	Positif	LP	LN
1	1 %	**57,1** %	7+11	7+11
	2 %	56,1 %	7+6	7+8
	3 %	41,3 %	7+5	7+1
2	1 %	54,0 %	7+11	7+18
	2 %	59,2 %	7+6	7+11
	3 %	58,9 %	7+3	7+4
3	1 %	59,8 %	7+11	7+32
	2 %	54,9 %	7+4	7+15
	3 %	57,8 %	7+5	7+6
4	1 %	64,7 %	7+16	7+30
	2 %	63,1 %	7+6	7+25
	3 %	63,2 %	7+3	7+9

Tableau 5. *Classification de 1 000 documents positifs et négatifs avec les mots appris et application d'AcroDef$_{IM3}$*

Adjectifs positifs supprimés	
Adjectifs	Nb occ.
Helpful	52
Inevitable	42
Attendant	4
Encouraging	2

Adjectifs négatifs supprimés			
Adjectifs	Nb occ.	Adjectifs	Nb occ.
Next	718	Current	37
Few	332	Unpleasant	22
Tricky	92	Unattractive	4
Legal	76	Unpopular	2
Environmental	2		

Tableau 6. *Listes des adjectifs positifs et négatifs éliminés par AcroDef$_{IM3}$ lors du premier apprentissage pour WS=1 et S=1 %*

rechercher de nouveaux adjectifs pertinents. Cette méthode est répétée tant que nous apprenons de nouveaux adjectifs.

Mots positifs appris	
Adjectifs	Nb d'occ.
Interesting	301
comic	215
Wonderful	165
Successful	105
Exciting	88

Mots négatifs appris	
Adjectifs	Nb d'occ.
Commercial	198
Dead	181
Terrible	113
Scary	110
Sick	40

Tableau 7. *Occurrences des adjectifs appris par le premier renforcement pour WS=1 et S=1 %*

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Les adjectifs appris lors de la première application de cette méthode sont illustrés dans le tableau 7. Les résultats obtenus en appliquant la procédure de classification sont présentés dans le tableau 8. Nous constatons que le pourcentage de documents positifs bien classés, s'est amélioré de 2.2 %, soit de 75.9 % à **78.1** %.

WS	S	Positif	LP	LN
1	1 %	**78,1** %	7+16	7+16

WS	S	Négatif	LP	LN
1	1 %	**54,9** %	7+16	7+16

Tableau 8. *Classification de 1 000 documents positifs et négatifs avec les mots appris et application d'AcroDef$_{IM3}$*

Les mots appris dans le deuxième renforcement sont ensuite considérés comme mots germes et ajoutés aux listes précédentes. Les nouveaux adjectifs appris sont présentés dans le tableau 9.

Adjectifs positifs appris	
Adjectifs	Nb d'occ.
special	282
entertaining	262
sweet	120

Adjectifs négatifs appris	
Adjectifs	Nb d'occ.
awful	109

Tableau 9. *Occurrences des adjectifs appris par second renforcement pour WS=1 et S=1 %*

Comme nous pouvons le constater dans le tableau 10, le résultat de classification sur le même jeu de test, s'est amélioré et est passé de 78.1 % à **78.7** % pour les documents positifs du jeu de test. Mais il a diminué légèrement pour les négatifs. La raison est que notre méthode de mesure du score est élémentaire et ponctuelle et se fonde sur le nombre d'occurrences des adjectifs. D'après les listes des adjectifs appris avec leur nombre d'occurrences, nous trouvons que le nombre d'occurrences des adjectifs positifs appris est notamment supérieur à celui des adjectifs négatifs ce qui influe sur notre méthode de classification.

WS	S	Positif	LP	LN
1	1 %	**78,7** %	7+19	7+17

WS	S	Négatif	LP	LN
1	1 %	**46,7** %	7+19	7+17

Tableau 10. *Classification de 1 000 documents positifs et négatifs avec les mots appris et application d'AcroDef$_{IM3}$*

Cependant, une nouvelle application de la méthode ne permet plus d'acquérir d'adjectifs. A cette étape, nous obtenons deux listes d'adjectifs (Cf. tableau 11) pertinents et discriminants pour le domaine *movies*.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Liste d'adjectifs positifs	
Adjectif	Adjectif
Good	Great
Nice	Funny
Excellent	Perfect
Superior	Beautiful
Positive	Worth
Correct	Major
Fortunate	Interesting
Hilarious	Comic
Happy	Wonderful
Important	Successful
Amazing	Exciting
Complete	Entertaining
Special	Sweet

Liste d'adjectifs négatifs	
Adjectifs	Adjectifs
Bad	Boring
Wrong	Different
Poor	Ridiculous
Nasty	Dull
Unfortunate	Silly
Negative	Expensive
Inferior	Huge
Certain	Dead
Dirty	Terrible
Social	Scary
Favorite	Sick
Awful	Commercial

Tableau 11. *Listes des adjectifs pour WS=1 et S=1 % pour le domaine « movies »*

WS	S	Positif	LP	LN
1	1 %	**82,6** %	7+19	7+17

WS	S	Négatif	LP	LN
1	1 %	**52,4** %	7+19	7+17

Tableau 12. *Classification de 1 000 documents positifs et négatifs avec les mots appris, application d'AcroDef$_{IM3}$ et négation*

De manière à améliorer notre méthodes de classification, nous avons intégré la prise en compte de la négation décrite dans la section 3.3. Nous constatons que les résultats de classification de 1 000 textes positifs se sont améliorés de 78.7 % à **82.6 %** et de 46.7 % à **52.4 %** pour les 1 000 textes négatifs comme le montre le tableau 12.

4.2. *Expérimentation sur la taille des jeux d'apprentissage*

Dans cette section, nous souhaitons répondre aux questions suivantes : à partir de combien de documents l'apprentissage devient robuste ? Quel est le nombre de documents suffisant pour chaque mot germe, à partir desquels les résultats convergent ? Pour répondre à ces questions, nous avons appliqué notre méthode d'apprentissage AMOD en faisant croître de 50 le nombre de documents à collecter jusqu'à atteindre la stabilité du nombre d'adjectifs appris.

Nous pouvons constater, dans la figure 11 qui illustre la relation entre la taille du corpus d'apprentissage et les adjectifs appris, que nous n'apprenons plus beaucoup de nouveaux adjectifs à partir de 2 800 documents (*i.e.* 200 documents pour chaque mot germe). Nous avons adopté la valeur de 300.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

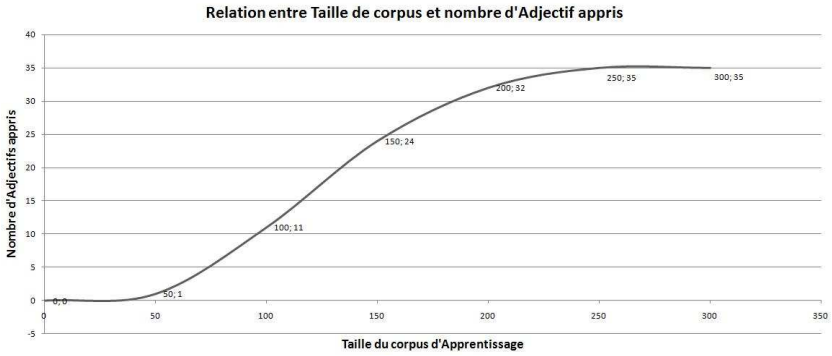

Figure 11. *Graphe de la relation entre nombre d'adjectifs appris et taille du corpus d'apprentissage*

4.3. *Comparaison avec une méthode de classification traditionnelle*

Dans cette section, notre objectif est de comparer les résultats obtenus avec une méthode de classification traditionnelle et l'approche AMOD. La comparaison est faite avec l'approche COPIVOTE (Plantié *et al.,* 2008) qui utilise un modèle d'apprentissage fondé sur le vote de plusieurs classifieurs. Les expérimentations ont été menées en considérant les mêmes corpus de test et d'apprentissage sur le cinéma que dans la section précédente.

Pour évaluer nos expérimentations, nous avons utilisé le FScore (Risbergen, 1979). Le FScore est donné par la formule suivante :

$$Fscore = \frac{(\beta^2+1)\times Précision\times Rappel}{(\beta^2)\times Précision+Rappel}$$

Les mesures d'évaluation *Précision* et *Rappel* sont données par les formules ci-après :

$$Rappel_i = \frac{Nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ appartenant\ à\ la\ classe\ i}$$

$$Précision_i = \frac{Nombre\ de\ documents\ correctement\ attribués\ à\ la\ classe\ i}{Nombre\ de\ documents\ attribués\ à\ la\ classe\ i}$$

Le Fscore donne un compromis entre Rappel et Précision, le paramètre β permet de régler les influences respectives de la Précision et du Rappel. Il est fixé à 1 pour accorder le même poids à ces deux mesures d'évaluation.

Nous pouvons remarquer dans le tableau 13 que notre approche est nettement supérieure dans le cas positif (71,73 % *vs.* 60,5 %), et qu'elle est meilleure dans le cas

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Docs :	Positifs	Négatifs
FScore COPIVOTE :	60,5 %	60,9 %
FScore AMOD :	**71,73 %**	**62,2 %**

Tableau 13. *Fscore du résultat de la classification de 1 000 documents de test négatifs et positifs avec* COPIVOTE *et* AMOD

négatif (62,2 % *vs.* 60,9 %). Alors que généralement la méthode COPIVOTE est particulièrement efficace dans le cadre de la classification de texte (*i.e.* étant basée sur une approche de vote, la méthode la plus efficace est sélectionnée), elle est pénalisée par le fait que les données de tests sont exprimées de manière différente de celles du jeu d'apprentissage.

4.4. *Application de l'approche* AMOD *à un autre domaine*

Pour vérifier que les choix effectués dans la spécification des paramètres de notre méthodologie restent valides et donc généralisables et que notre approche peut être utilisée pour déterminer automatiquement les adjectifs d'un autre domaine, nous l'avons appliquée dans le domaine automobile avec comme mot-clé « *car* ». De manière à valider les connaissance acquises, nous utilisons 40 documents positifs extraits de *www.epinions.com* pour les tests. Le résultat de la classification en ne considérant que les mots germes initiaux (*i.e.* sans appliquer la méthode d'extraction des adjectifs), est de **57.5 %**, comme le montre le tableau 14.

WS	S	Positif	LP	LN
1	1 %	**57,5** %	7+0	7+0

Tableau 14. *Classification de 40 documents positifs avec les mots germes*

L'application de l'approche AMOD, avec comme valeurs de paramètres WS=1 et support = 1 %, a permis d'apprendre les adjectifs illustrés dans le tableau 15. Nous pouvons constater que le nouveau résultat est passé de 57.5 % à **87.5 %** comme le montre le tableau 16.

Les adjectifs appris lors d'un nouveau renforcement sont illustrés dans le tableau 17.

Les résultats obtenus en appliquant la procédure de classification sont précisés dans le tableau 18, nous constatons que le pourcentage de documents positifs bien classés, s'est amélioré de 7.5 %, soit de 87.5 % à **92.5 %**.

Le résultat de la classification avec un premier renforcement est décrit dans le tableau 19. Nous pouvons constater que nous sommes passés de 92.5 % à **95 %**. Le

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Adjectifs positifs appris	
Adjectifs	Adjectifs
Comfortable	Great
Professional	Full
Popular	Fabulous
Powerful	Economical
Luxurious	Quiet
Secured	Strong
Several	

Adjectifs négatifs appris	
Adjectifs	Nb d'occ.
Unsecured	Expensive
Horrible	Ugly
Uncomfortable	Lack

Tableau 15. *Listes des adjectifs appris*

WS	S	Positif	LP	LN
1	1 %	**87.5** %	7+13	7+6

Tableau 16. *Classification de 40 documents positifs avec les mots appris et application d'AcroDef*$_{IM3}$

Adjectifs positifs appris	
Adjectifs	Adjectifs
Practical	Maximum
Lovely	First
Successful	Active
Amazing	Super

Adjectifs négatifs appris	
Adjectifs	Nb d'occ.
Heavy	Dangerous
Weird	Boring

Tableau 17. *Adjectifs appris après un renforcement*

WS	S	Positif	LP	LN
1	1 %	**92.5** %	7+21	7+10

Tableau 18. *Classification de 40 documents positifs avec les mots appris et application d'AcroDef*$_{IM3}$

tableau 20 décrit les adjectifs appris pour le domaine voiture. Par rapport à la section précédente, les deux jeux de données concernent des blogs et sont donc exprimées de la même manière. Nous montrons ainsi que notre approche obtient également de meilleurs résultats pour des jeux de données similaires.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

WS	S	Positif	LP	LN
1	1 %	**95** %	7+26	7+10

Tableau 19. *Classification de 40 documents positifs avec les mots appris, application d'AcroDef$_{IM3}$ et négations*

Liste des adjectifs positifs	
Adjectifs	Adjectifs
Good	Comfortable
Nice	Powerful
Excellent	Fabulous
Superior	Economical
Positive	Quiet
Correct	Strong
Fortunate	Several
Professional	Lovely
popular	Successful
Luxurious	Amazing
Secured	Maximum
Great	First
Full	Active
Efficient	Beautiful
hard	Wonderful
Fast	Practical

Liste des adjectifs négatifs	
Adjectifs	Adjectifs
Bad	Unsecured
Wrong	Uncomfortable
Poor	Expensive
Nasty	Ugly
Unfortunate	Luck
Negative	Heavy
Inferior	Dangerous
Horrible	Weird
Boring	

Tableau 20. *Listes des adjectifs pour WS=1 et S=1 % pour le domaine voiture*

4.5. *Discussion*

Comme nous avons pu le constater lors de nos expérimentations, le fait de se concentrer sur les adjectifs a permis de bien détecter l'opinion exprimée dans les textes. De même, l'utilisation des formes négatives a amélioré considérablement la classification. Nous avons pu constater que notre approche était suffisamment automatique pour pouvoir être appliquée dans différents domaines et ainsi extraire les adjectifs significatifs de ces derniers. Revenons à présent sur les corpus utilisés dans le cadre du cinéma, nous avons également pu montrer que notre approche était telle qu'elle surpassait les approches de classification traditionnelle. Cette différence montre bien que le fait d'apprendre les adjectifs spécifiques apporte un réel plus dans le cadre de la détection d'opinions.

En ce qui concerne les travaux présentés dans l'état de l'art (Turney, 2002; Turney *et al.*, 2002; Turney *et al.*, 2003) utilisent également le web comme corpus d'apprentissage, mais ils extraient l'orientation sémantique des mots, à partir de tous les documents existants, sans prendre en compte la dépendance entre les documents et le

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

contexte. En outre, ils infèrent l'orientation sémantique à partir du nombre de documents retournés par un moteur de recherche (*i.e.* un document est retenu s'il contient le mot cherché et le mot germe, sans tenir compte de la distance entre les deux).

Les limites auxquelles nous avons fait face sont dues au moteur de recherche blogsearch.google.fr. Il n'était pas possible de proposer à Google une requête de plus de 10 mots-clés, ce qui influe sur la qualité des documents retournés. Donc, nous avons capté du bruit dans les documents collectés, et ce bruit est lié à la présence des mots germes à ignorer.

Par contre, notre méthode d'extraction de l'orientation sémantique était plus décisive et plus ciblée que les approches existantes, puisque nous cherchions, dans des fenêtres de plusieurs tailles la corrélation entre les mots germes et tous les adjectifs qui pourraient être utilisés dans un contexte donné. Le fait d'appliquer un filtre $AcroDef_{IM3}$ sur les résultats a également permis d'améliorer la qualité des adjectifs pertinents au domaine traité.

De manière à valider nos connaissances acquises, nous avons utilisé nos adjectifs pour classer des documents selon leur orientation sémantique. Les travaux de (Turney, 2002; Turney *et al.*, 2002; Turney *et al.*, 2003; Taboada *et al.*, 2006; Voll *et al.*, 2007), s'appuient sur des méthodes classiques, qui calculent le nombre d'occurrences des mots afin de prédire l'orientation sémantique, sans aucun traitement syntaxique du texte. En utilisant une analyse morpho-syntaxique pour rechercher les différentes forment de négation qui influent sur l'opinion exprimée nous avons également pu améliorer la qualité de la classification.

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 expérimentations menées sur des jeux de données issus de l'apprentissage (blogs *vs.* reviews de cinéma et de voiture) ont montré que *via* notre approche nous étions capables d'apprendre les adjectifs pertinents et que surtout celle-ci était généralisable à différents domaines et était moins sensible à la manière d'exprimer des opinions que les approches de classification traditionnelles.

Les perspectives à ce travail sont nombreuses. Tout d'abord, notre méthode d'apprentissage dépend fortement de la qualité des documents constituant le corpus d'apprentissage. Notre corpus d'apprentissage est pour l'instant créé à partir de requêtes spécifiques exécutées sur blogsearch.google.fr. L'originalité de notre requête tient au fait de concilier un ensemble de mots à inclure et en même temps un autre ensemble de mots à éliminer. Lors de nos expérimentations, nous avons remarqué toutefois la limite de blogsearch.google.frsearch sur ce point (*i.e.* la qualité des documents rendus, et le nombre des mots-clés associés à chaque requête). Une perspective serait d'étendre notre méthode de création de corpus d'apprentissage en prétraitant les documents collectés par Google (*i.e.* en supprimant automatiquement les documents ne

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

respectant pas véritablement la requête) afin d'avoir un corpus d'apprentissage moins bruité.

Notre méthode d'extraction se fonde sur la génération des règles d'association entre les mots germes et les adjectifs. Lors de nos expérimentations nous avons principalement considéré, *via* les règles, des corrélations entre les différents adjectifs. Une autre perspective est d'étudier si l'ordre entre les adjectifs peut avoir un impact sur la classification. Traditionnellement, les approches de fouille textes considèrent plutôt les n-grammes pour considérer l'ordre entre les mots ou les caractères. L'avantage des n-grammes est bien entendu de retrouver des adjectifs qui sont très proches (*i.e.* en fonction de la valeur de *n*). Le défaut de ces approches dans le contexte de l'opinion mining est qu'elles nécessitent que les adjectifs soient très proches afin de les repérer. Notre idée est d'étendre l'approche en utilisant la notion de motifs séquentiels et permettre ainsi d'extraire des adjectifs qui sont proches sans être consécutifs. Dans ce cas, nous serions capables de reconnaître non seulement que des adjectifs utilisés pour exprimer une opinion sont ordonnés mais en plus qu'ils interviennent dans une même fenêtre. Cette notion de motifs séquentiels peut également être utilisée pour extraire les formes syntaxiques d'expression d'opinion. Dans ce cas, l'objectif est plutôt de s'intéresser au style d'expression des opinions (e.g. une opinion négative contient un verbe suivi par un adjectif suivi par un adverbe).

Une autre perspective de ce travail est d'étendre la méthode d'extraction à d'autres catégories que les adjectifs. En fait, la génération des règles d'associations entre les mots germes et tous les termes du corpus d'apprentissage devrait nous permettre d'extraire non seulement des adjectifs mais aussi des verbes ou des adverbes généralement discriminants pour un domaine donné.

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.

Church K., Hanks P., « Word association norms, mutual information and lexicography », *Proceedings of the 27th Annual Conference of the ACL*, New Brunswick, NJ, p. 76-83, 1989.

Daille B., « Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filltres linguistiques », *Thèse de Doctorat, Université Paris VII, France*, 1994.

Esuli A., Sebastiani F., « Determining the Semantic Orientation of Terms through Gloss Analysis », *In Proceedings of CIKM-05, the 14th ACM international Conference on Information and Knowledge Management*, Bremen, Germany, 2005.

Grouin C., Berthelin J.-B., Ayari S. E., Heitz T., Hurault-Plantet M., Jardino M., Khalis Z., Lastes M., « Présentation de DEFT'07 (Défi Fouille de Textes) », *Proceedings of the DEFT'07 workshop, Plate-forme AFIA, Grenoble, France*, 2007.

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

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Hatzivassiloglou V., Wiebe J., « Effects of adjective orientation and gradability on sentence subjectivity », *Actes de International Conference on Computational Linguistics (COLING'00)*, Saarbrücken, Germany, 2000.

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

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

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

Plantié M., Extraction automatique de connaissances pour la décision multicritère, PhD thesis, École Nationale Supérieure des Mines de Saint Etienne et de l'Université Jean Monnet de Saint Etienne, Nîmes, 2006.

Plantié M., Roche M., Dray G., Poncelet P., « Is a voting approach accurate for opinion mining ? », *Proceedings of the 10th International Conference on Data Warehousing and Knowledge Discovery (DaWaK '08)*, Torino Italy, 2008.

Quinlan J., « Induction of decision trees », *Mach learn*, p. 81-106, 1986.

Risbergen V., « Information Retrieval, 2nd edition », *Butterworths*, London, 1979.

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 », MIT Press, Cambridge, MA, 1966.

Strapparava C., Valitutti A., « WordNet-Affect : and Affective Extension of WordNet », *In Proceedings of LREC-04, the 4th Conference on Language Resources and Evaluation*, Lisbon, Portugal, 2004.

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*, Paris, p. 417-424, 2002.

Turney P., Littman M., « Unsupervised learning of semantic orientation from a hundred-billion-word corpus », *National Research Council of Canada*, 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.

Yang H., Si L., Callan J., « Knowledge Transfer and Opinion Detection in the TREC2006 Blog Track », *Notebook of Text REtrieval Conference*, 2006.

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com

Cet article des Editions Lavoisier est disponible en acces libre et gratuit sur dn.revuesonline.com