VERS UNE CARACTÉRISATION AUTOMATIQUE DE CRITÈRES POUR L'OPINION-MINING

Benjamin Duthil *et al.*

Lavoisier I *Les Cahiers du numérique*

2011/2 - Vol. 7
pages 41 à 62

ISSN 1622-1494

Article disponible en ligne à l'adresse:

---

http://www.cairn.info/revue-les-cahiers-du-numerique-2011-2-page-41.htm

---

Pour citer cet article :

Duthil Benjamin*et al.*, « Vers une caractérisation automatique de critères pour l'opinion-mining »,
Les Cahiers du numérique, 2011/2 Vol. 7, p. 41-62.

Distribution électronique Cairn.info pour Lavoisier.

VERS UNE CARACTÉRISATION AUTOMATIQUE DE CRITÈRES POUR L'OPINION-MINING

BENJAMIN DUTHIL
FRANÇOIS TROUSSET
GÉRARD DRAY
JACKY MONTMAIN
PASCAL PONCELET

Les technologies de l'information et le succès des services associés (e.g. blogs, forums...) ont ouvert la voie à un mode d'expression massive d'opinions sur les sujets les plus variés (sites de e-commerces, films, etc.). Récemment, de nouvelles techniques de détection automatique d'opinion (opinion-mining) proposent des analyses statistiques des avis exprimés afin de dégager une tendance globale des opinions recueillies sur une entité évaluée. Néanmoins, une analyse plus fine montre que les arguments des internautes relèvent de critères de jugement distincts. Les approches d'opinion-mining traitent peu de cet aspect multicritère pourtant riche en informations. Dans cet article, nous proposons de caractériser automatiquement les segments de textes relevant d'un critère donné sur un corpus de critiques. À partir d'un ensemble restreint de mots-clés associés à un critère, notre approche construit automatiquement une base d'apprentissage. Des expériences menées sur des jeux de données réelles illustrent l'efficacité du processus.

DOI:10.3166/LCN.7.2.41-62

1. Introduction

Avec le développement du web, de plus en plus de documents textuels sont disponibles et de plus en plus d'outils permettent de pouvoir rechercher de l'information pertinente. Connaître l'opinion des personnes sur un produit, rechercher et classer des documents, indexer de manière automatique des documents sont des problématiques d'actualité. Par exemple, dans le cas de l'opinion de cinéphiles, de nombreux outils sont disponibles [1] pour connaître l'avis général des spectateurs sur un film. Cependant l'information trouvée sur le web ne restitue pas toujours toute la richesse sémantique d'une critique. Considérons par exemple la Figure 1 qui donne l'opinion d'une personne sur le film *Avatar*. Nous pouvons constater que la note attribuée par le cinéphile est de 9,5/10. Cette note peut s'expliquer de la façon suivante : le scénario du film devrait lui valoir un score décevant mais la remarquable mise en scène hyperréaliste du monde imaginaire de James Cameron semble être le principal centre d'intérêt du critique et contribue donc le plus significativement à la bonne évaluation du film. Le score agrégé masque la divergence sur les critères mise en scène et scénario et restitue donc mal la richesse sémantique du texte.

Ce constat n'est pas spécifique à l'analyse de critiques cinématographiques. Cette problématique se retrouve, par exemple, dans le domaine de la politique (Thomas *et al.*, 2006), le e-commerce (Castro-Schez *et al.*, 2011), les systèmes de recommandations (Garcia *et al.*, 2011).

Figure 1. Un exemple d'opinion

1. e.g. http://www.premiere.fr/Cinema/Critique-Film

Dans cet article, nous proposons une approche qui réponde à cette problématique : nous extrayons automatiquement depuis le web des textes se rapportant à un ensemble de critères prédéfinis afin de restituer aux internautes des critiques plus pertinentes et mieux ciblées sur leurs critères de choix personnels. On désigne désormais par critique tout extrait de texte rapportant une opinion relativement à un critère. L'extraction automatique de critiques nécessite donc de caractériser chacun des critères. Cette caractérisation est traditionnellement réalisée dans les approches existantes (e.g. Mindserver Categorization, Thunderstone...) à l'aide d'une classification supervisée. Cette dernière, à partir d'un ensemble d'apprentissage annoté, permet d'apprendre les descripteurs utiles à la classification des critères. Cependant, dans le contexte du web, la constitution d'un corpus expertisé est très coûteuse, voire décourageante. En effet, de par la diversité des documents (e.g. blogs, forum, dépêches journalistiques), les niveaux de langage varient significativement d'un support à l'autre, ce qui multiplie le nombre de descripteurs à identifier. Reprenons l'exemple du critère scénario. Alors que nous pourrons trouver sur un forum la phrase : « le scénard est élémentaire », elle sera exprimée de manière différente dans une dépêche. De la même manière, si nous considérons un critère acteur, la manière d'orthographier les noms peut varier d'un individu à l'autre. Ainsi le nom de l'acteur principal du film *Avatar,* Sam Worthington, sera orthographié Wortington ou Wortingthon. De plus, le volume de données à traiter rend la tâche d'annotation manuelle ardue, voire impossible. Ce dernier constat met en évidence l'intérêt de construire un corpus d'apprentissage automatiquement ou du moins avec une intervention minimale de l'homme.

Les principales contributions de cet article sont les suivantes. Tout d'abord, plutôt que d'essayer d'associer une note globale et l'opinion générale associée à un document, nous proposons une analyse plus fine en décomposant cette opinion en avis relatifs à un ensemble de critères caractéristiques du domaine (dans l'exemple précédent, il s'agit de mise en scène et scénario pour le domaine du cinéma). En outre, nous proposons une méthode de construction automatique d'un corpus d'apprentissage en utilisant une expertise minimale (sélection des critères pertinents et de quelques mots-clés associés).

Basée sur une approche statistique d'apprentissage, nous proposons de construire un lexique de descripteurs caractérisant chacun des critères à évaluer. Ces lexiques sont utilisés par la suite pour identifier de manière automatique les parties de textes correspondant à des critères dans des documents. Enfin, nous montrons qu'il est possible d'intégrer un processus d'opinion-mining pour évaluer automatiquement l'opinion relative à chacun des critères.

L'article est organisé de la manière suivante. Dans la section 2, nous décrivons dans un premier temps le fonctionnement général de l'approche puis,

de manière détaillée, les différentes étapes. La section 3 présente la méthode d'extraction thématique utilisée par le processus d'opinion-mining décrit dans la section 4. Les expérimentations menées sont décrites dans la section 5. Un état de l'art est présenté dans la section 6 et nous concluons dans la section 7 en proposant quelques perspectives.

2. L'approche

Dans cette section, nous présentons tout d'abord une brève description du processus global. Nous décrivons, par la suite, les étapes d'acquisition de documents, d'apprentissage des mots caractéristiques des critères (lexique), d'extraction thématique dans des données textuelles et de détection d'opinion.

2.1. Présentation générale

Dans cet article, nous avons choisi le cinéma (movie) comme domaine d'application et nous nous intéressons à deux critères : acteur (actor) et scénario (scenario). L'architecture générale de l'approche est décrite dans la Figure 2.

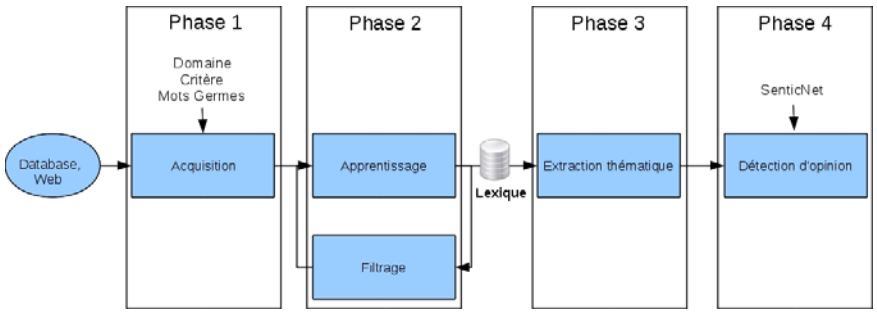

Figure 2. Architecture générale

– Phase 1. L'objectif de cette phase est de construire automatiquement un corpus d'apprentissage pour chaque critère. Le processus débute de la manière suivante. Après avoir précisé un domaine et des critères, l'utilisateur spécifie, pour chacun des critères, un ensemble restreint de mots germes caractéristiques de chaque critère. Par exemple, considérons les germes suivants associés aux critères : pour *scenario*, « adaptation, narrative, original, scenario, scriptwriter,

story, synopsis », et pour *actor* « acting, actor, casting, character, interpretation, role, star ».

Pour chaque germe d'un critère, le système recherche, en interrogeant un moteur de recherche, des documents du domaine possédant au moins une occurrence du germe pour le critère considéré. L'objectif est de pouvoir rechercher, par la suite, dans ces documents les mots corrélés à ces mots germes : les germes permettent d'engendrer le lexique propre au critère. L'union de tous les textes associés aux mots germes d'un critère constitue la classe du critère. La classe est le premier élément constitutif du corpus associé à un critère.

Exemple 1. La requête suivante illustre la recherche du germe casting pour le critère actor à l'aide de Google : « +movie +casting -acting -actor -character - interpretation -role -star ». Les symboles + (resp. -) indiquent que ces mots doivent se retrouver (resp. ne pas se retrouver) dans le document retourné.

De la même manière, un second ensemble de documents, appelé anti-classe du critère, est constitué en recherchant des documents du domaine qui ne contiennent aucun des germes du critère. Cet ensemble constitue la seconde partie du corpus et a pour objectif de mieux caractériser les éléments du critère : un terme du domaine, fréquent dans la classe et dans l'anti-classe, n'est pas caractéristique du critère. Le corpus sera donc composé des deux ensembles de documents : classe et anti-classe. Les textes de chaque corpus font l'objet ensuite d'une analyse morpho-syntaxique et d'une lemmatisation.

Exemple 2. Le texte « Le rôle de cet acteur » est transformé de la manière suivante : le (DET :ART) rôle (NOM) de (PRP) ce (PRO :DEM) acteur (NOM).

– Phase 2. À partir des textes filtrés en phase 1, nous cherchons à déterminer les mots représentatifs (resp. non représentatifs) de chacun des critères. Pour cela, nous étudions la fréquence des mots fortement corrélés aux germes dans les textes du corpus relatif au critère. Nous considérons l'hypothèse suivante : plus les mots fréquents sont proches d'un germe, plus ils ont de chances de caractériser le critère de manière à ne se focaliser que sur les mots proches des mots germes. À l'issue de l'analyse nous obtenons un ensemble de mots proches des germes avec leur fréquence d'apparition dans les documents.

On distingue quatre types de mots :

– les très fréquents dans la classe et peu fréquents dans l'anti-classe ;

– les très fréquents dans l'anti-classe et peu fréquents dans la classe ;

– les très fréquents dans les deux classes ;

– les peu fréquents dans les deux classes ;

Dans la troisième catégorie, les mots ne sont pas caractéristiques du critère. Pour la dernière catégorie, il n'est pas possible, à partir des seuls documents du corpus de savoir si les mots sont caractéristiques ou non : le corpus devra être enrichi (cf. section 2.2.4). Le lexique relatif à un critère sera le résultat d'un filtrage sélectif sur les catégories un, deux et quatre (cf. section 2.2.5).

– Phase 3. À partir de ces lexiques, nous proposons une technique permettant d'isoler les segments de texte qui sont associés à chacun des critères.

– Phase 4. En appliquant un processus d'opinion-mining sur les segments identifiés dans la phase 3, nous calculons une opinion relative à un critère. Ce processus utilise *SenticNet* (Cambria *et al.*, 2010) pour détecter et attribuer un score à chaque mot porteur d'opinion.

2.2. Caractérisation des critères

2.2.1. Acquisition

Comme nous l'avons vu dans la section précédente, la première phase consiste à acquérir et à prétraiter des documents issus du web via un moteur de recherche (cf. exemple 1).

L'utilisateur fournit n mots germes caractéristiques d'un critère C. Pour chaque mot germe g d'un critère C le système recherche 300 documents contenant à la fois ce mot germe et le mot D, i.e. le nom du domaine. L'ensemble des documents ainsi obtenus pour chacun des mots germes du critère constitue la classe du critère C. De la même manière, nous recherchons 300 documents du domaine D mais ne contenant aucun des mots germes du critère C. Ceci constitue l'anti-classe du critère C. La classe d'un critère C est alors composée de n * 300 documents (où n est le nombre de mots germes). L'anti-classe comprend 300 documents. Les balises HTML, les publicités, etc. sont ensuite éliminées des documents du corpus de C. Ces documents sont enfin transformés à l'aide d'un analyseur morphosyntaxique (cf. exemple 2).

2.2.2. Apprentissage des mots significatifs

Il s'agit de rechercher les mots fortement corrélés aux mots germes, i.e. les mots qui apparaissent fréquemment très proches des germes. Pour chaque document *t*, nous utilisons une fenêtre *F* centrée sur les germes. Cette dernière est définie plus formellement de la manière suivante :

$$F(g, sz, t) = \{m \in t \,/\, d^{t}_{NC}(g, m) \leq sz\} \qquad (1.1)$$

où g correspond au germe, sz représente la taille de la fenêtre et $d^t_{NC}(g,m)$ est la distance correspondant au nombre de noms communs séparant un mot m de g. Dans le choix de sz, nous nous focalisons sur les noms communs qui sont reconnus comme mots porteurs de sens dans un texte (Kleiber, 1996).

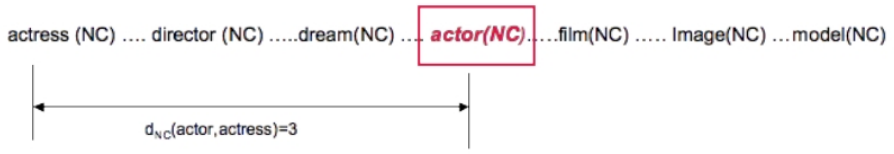

Figure 3. Exemple d'une fenêtre de taille 3

La Figure 3 illustre un exemple de fenêtre de taille 3, i.e. il y a 3 noms communs à gauche (actress, director, dream) et 3 noms communs à droite (film, image, model) du germe *actor*.

Dans une fenêtre, plus un mot est voisin du germe, plus il est considéré comme caractéristique du critère : on dit qu'il est influencé par le germe. L'influence $I(m,g,sz,t)$ est définie de la manière suivante :

$$I(m,g,sz,t) = \begin{cases} 0 \; si \; m \notin F(g,sz,t) \\ h(d^t_*(m,g)) \; si \; m \in F(g,sz,t) \end{cases} \quad (1.2)$$

où $d^t_*(m,g)$ est la distance définie par le nombre de mots entre g et m, sans considération de la nature grammaticale (noté *), h une fonction de lissage qui pourra, par exemple, correspondre à deux demi-gaussiennes centrées sur g. Soient l et r les mots les plus éloignés à gauche et à droite de g dans la fenêtre, l'influence est alors donnée pour un filtrage gaussien par $gauss\left(\frac{d^t_*(g,M)}{d^t_*(g,l)},\mu,\sigma\right)$ pour un mot à gauche de g et par $gauss\left(\frac{d^t_*(g,M)}{d^t_*(g,r)},\mu,\sigma\right)$ pour un mot à droite de g.

2.2.3. Représentativité

Nous pouvons maintenant définir pour chaque mot sa représentativité respectivement dans la *classe* X et dans l'*anti-classe* $\overline{X}$ du critère C. On appelle $O(M,T)$ les occurrences du mot M dans un texte T. Pour un mot M, sa représentativité dans la *classe* est calculée de la manière suivante (avec S l'ensemble des mots germes) :

$$X(M,sz) = \sum_{g \in S} \sum_{t \in T(g)} \sum_{\gamma \in O(g,t)} \sum_{m \in O(M,t)} I(m,g,sz,t) \tag{1.3}$$

Ce qui correspond à cumuler l'influence de tous les mots germes de g sur le mot M pour tous les textes de la *classe*. Sa représentativité dans l'*anti-classe* est définie par :

$$\overline{X}(M,sz) = \sum_{t \in anti-classe} \sum_{m \in O(M,t)} I(m,D,sz,t) \tag{1.4}$$

Ce qui correspond à cumuler l'influence du domaine D sur le mot M pour tous les documents de l'*anti-classe*.

	X	$\overline{X}$
film	1080	460
actress	170	0
theater	0	370
poster	700	700
Matt Vaughn	1	1
Sam Worthington	1	1
story	100	120

Tableau 1. Exemple de mots avec leurs fréquences respectives dans la classe et dans l'anti-classe pour le critère actor

Le Tableau 1 illustre un exemple de représentativité des mots dans la *classe* et l'*anti-classe* pour le critère *actor*. Nous retrouvons les catégories de mots comme indiqué dans la section 2.1. Tout d'abord, nous constatons que le mot « poster » apparaît fréquemment dans les deux classes : il n'est pas discriminant pour le critère *actor*, il est éliminé. Le mot « film », par contre, apparaît beaucoup plus

fréquemment dans la classe que dans l'anti-classe et est donc considéré comme caractéristique du critère *actor*. En revanche, nous ne pouvons pas déduire d'appartenance sur les mots « Matt Vaughn » [2] et « Sam Wothington » puisqu'ils sont très peu fréquents dans les deux classes. La Figure 4 donne une représentation graphique des catégories de mots de la section 2.1.

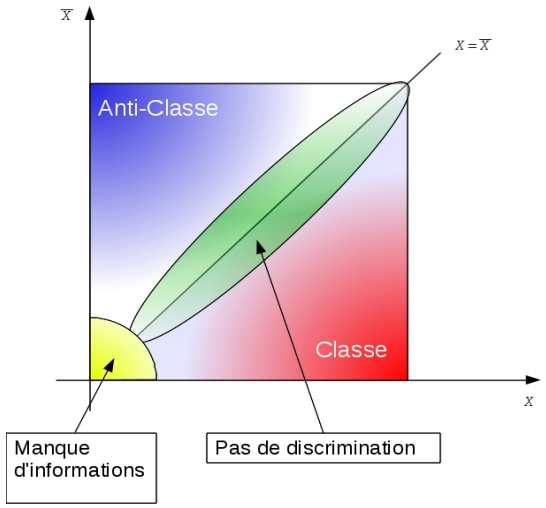

Figure 4. Représentation des différentes catégories de mots de la section 2.1

2.2.4. Validation des mots candidats et enrichissement du corpus

Les mots peu fréquents dans la classe et l'anti-classe sont appelés mots candidats : il manque de l'information pour qu'ils puissent être affectés à une classe. Nous enrichissons donc le corpus en recherchant sur le web de nouveaux documents : on lance n requêtes composées du domaine D, d'un des n mots germes et du mot candidat. La seule différence par rapport à la section 2.2.1 réside en l'ajout aux requêtes du mot candidat. X et $\overline{X}$ sont mises à jour avec ce corpus enrichi.

2. Par souci de simplification, nous considérons que « Matt Vaughn » constitue un seul mot.

2.2.5. Discrimination

En fonction de la représentativité d'un mot dans la *classe* et dans *l'anti-classe* (équations 1.3 et 1.4), nous établissons un score pour ce mot pour une taille de fenêtre donnée en utilisant la fonction de discrimination suivante :

$$Sc(M,sz) = \frac{(X(M,sz) - \overline{X}(M,sz))^3}{(X(M,sz) + \overline{X}(M,sz))^2} \quad (1.5)$$

La puissance cubique du numérateur permet une discrimination signée : les mots fréquents non représentatifs du critère (forte présence dans l'anti-classe) auront des scores négatifs, les mots représentatifs (forte présence dans la classe et faible présence dans l'anti-classe) auront quant à eux des scores positifs. La puissance carré du dénominateur permet de normaliser ce score. On peut alors construire le lexique propre à un critère. Il est composé de la liste des mots scorés pour ce critère (Tableau 2).

	Score
film	1040
actress	450
theater	-460
Sam Worthington	2640

Tableau 2. Extrait de lexique pour le critère actor

Le lexique peut maintenant être utilisé dans la phase 3 d'extraction thématique (section 3).

3. Extraction thématique

Dans cette section, nous illustrons comment utiliser le lexique dans un contexte d'extraction thématique dans des données textuelles. Nous nous concentrons sur l'identification des parties d'un document relatives à un critère C. Pour un texte t, nous introduisons une notion de fenêtre glissante de taille sz (c.f. section 2.2.2) successivement centrée sur chaque occurrence d'un nom commun dans le texte t. À partir du lexique de C, un score est calculé pour chacune des fenêtres f de la manière suivante :

$$Score(f) = \sum_{M \in f} Sc(M, sz) \tag{1.6}$$

Pour un texte t donné, nous considérons qu'une fenêtre est en rapport avec le critère, lorsque son score est supérieur à une valeur seuil th. Pour des contraintes de longueur, l'analyse de sensibilité qui permet de fixer la valeur de th ne peut être décrite dans cet article. Nous en donnons simplement le principe. L'idée consiste à analyser le nombre de mots susceptibles d'être liés au critère en fonction de la valeur du seuil. Les variations du nombre de mots retenus en fonction du seuil appliqué sont très lentes, sauf pour quelques valeurs remarquables (Figure 5). Celles-ci correspondent à des niveaux distincts de granularité liés à la structure thématique du document et au point de vue de l'utilisateur. Notre algorithme choisit automatiquement parmi chacun de ces points caractéristiques, celui offrant le plus large point de vue, i.e, celui qui maximise le nombre de mots retenus (200 dans l'exemple Figure 5).

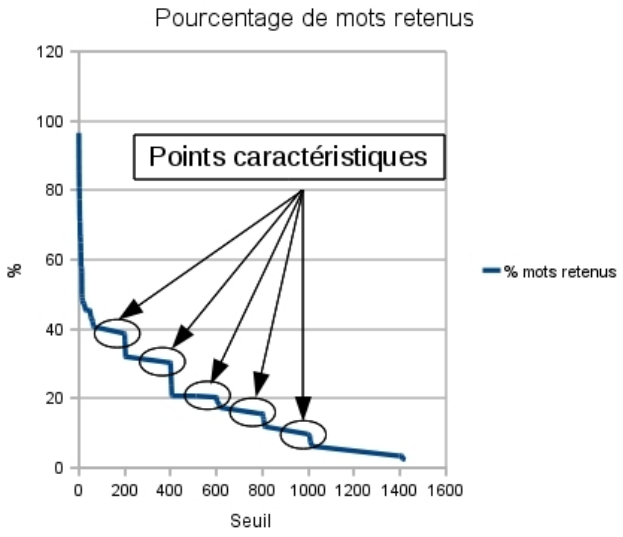

Figure 5. Pourcentage de mots conservés en fonction de la valeur du seuil choisi

4. Détection d'opinion

La problématique de détection d'opinion ou de sentiments dans des données textuelles est actuellement une des préoccupations principales des chercheurs dans de nombreux domaines comme le marketing, la prévision des

marchés financiers, l'analyse d'avis... (c.f. section 6). Les approches existantes pour l'extraction automatique d'opinion reposent sur l'identification des parties de texte exprimant des jugements de valeur. La majorité des approches se basent sur la détection de termes explicitant directement une appréciation, par exemple, les termes : bon, agréable, excellent, mauvais, méchant, brutal... Malheureusement, on se rend compte qu'une extraction d'opinion avec ces seuls termes explicites n'est pas suffisante pour assurer un résultat satisfaisant : l'expression d'opinion est au moins en partie propre au contexte (« Pierre est tout de même un *bon* gars » et « Orsi est un *bon* restaurant de Lyon »), au domaine (« l'autofocus de cet appareil photo est *parfaitement silencieux* » et « ce député est resté *parfaitement silencieux* tout au long de l'assemblée »), ce qui rend les approches purement syntaxiques inefficaces. Il est donc essentiel d'identifier les expressions et concepts propres au contexte ou à un domaine d'utilisation.

SenticNet (Cambria *et al.,* 2010) propose une collection de concepts polarisés (positifs/négatifs) qui constitue un réseau sémantique. Un score ($\in [-1,1]$: <0 pour l'expression d'une opinion plutôt négative et >0 pour l'expression d'une opinion plutôt positive) est attribué à chaque concept. Contrairement à la majorité des autres outils qui à chaque sens d'un concept attribuent un score différent (*http://sentiwordnet.isti.cnr.it/*), *SenticNet* propose une agrégation des différents sens, ce qui permet d'avoir un score unique adapté à tout contexte et ainsi d'éviter les redondances dans la collection de qualificatifs.

En combinant l'extraction de texte relative à un critère (c.f. section 3) à un processus d'opinion-mining, nous sommes en mesure de proposer une chaîne d'extraction d'opinion complète. La méthode consiste à détecter l'opinion sur les extraits identifiés par notre approche en utilisant *SenticNet*. Le processus se décompose en trois étapes :

– L'étape 1 identifie les éléments (mots et expressions) porteurs d'opinion sur un critère C pour chaque extrait identifié comme relevant du critère C par notre méthode d'extraction thématique.

– L'étape 2 d'évaluation attribue à chaque élément porteur d'opinion un score, puis calcule un score moyen pour chaque phrase.

– L'étape 3 décisionnelle agrège les scores obtenus pour chaque phrase d'un extrait de texte et établit un avis global pour celui-ci.

Nous sommes en mesure grâce à ce processus d'établir un score d'opinion à chaque extrait de texte identifié comme relatif à un critère. Il devient alors facile de classer un document en fonction de l'opinion qu'il contient en effectuant une simple moyenne des scores obtenus pour chaque extrait.

5. Expérimentations

De manière à étudier les performances de l'approche, nous avons réalisé de nombreuses expérimentations. Ces dernières ont été menées en utilisant le moteur de recherche *Google* et *TreeTagger* (Schmid, 1994) comme lemmatiseur et analyseur morphosyntaxique. Nous avons validé notre approche pour deux domaines : le cinéma (*movie*), avec comme critères acteur (*actor*) et scénario (*scenario*), et la restauration (*restaurant*), avec comme critères le service (*service*) et la propreté (*cleanliness*). Les mots germes des quatre critères sont :

– movie

- actor : acting, actor, casting, character, interpretation, role, star;

- scenario : adaptation, narrative, original, scenario, screenplay, scriptwriter, story, synopsis.

– restaurant

- service : server, tip, reservation, usability, waiter, service;

- cleanliness : health, appearance, cleanliness, restroom, ventilation, utensils.

Nous avons utilisé un corpus de test de critiques, en langue anglaise, provenant de différentes sources : blogs, critiques journalistiques..., étiqueté manuellement par des experts pour les deux domaines (700 phrases étiquetées pour chaque critère des domaines) avec en moyenne une cinquantaine d'extraits correspondant aux critères. Nos expérimentations se décomposent en deux parties. La première montre les performances du système pour l'extraction thématique. La seconde synthétise les performances du système sur la détection d'opinion sur les extraits de textes précédemment identifiés par le système.

5.1. Performances du système pour l'extraction thématique

Nous proposons ici de montrer les performances du système dans un contexte d'extraction thématique. Pour cela, nous utilisons comme jeux de test, les corpus annotés pour chacun des deux domaines. Les indicateurs classiques de validation utilisés sont le rappel, la précision et le FScore. Ils sont définis comme suit :

– $rappel = \frac{nombre\ de\ mots\ correctement\ attribués\ au\ critère}{nombre\ de\ mots\ appartenant\ au\ critère}$

– $précision = \frac{nombre\ de\ mots\ correctement\ attribués\ au\ critère}{nombre\ de\ mots\ attribués\ au\ critère}$

– $FScore = 2 * \frac{Rappel * Precision}{Rappel + Precision}$

Performances du système pour l'extraction thématique				
	domaine *cinéma*		domaine *restauration*	
	acteur	*scénario*	*propreté*	*service*
FScore	0,89	0,77	0,67	0,74
précision	0,87	0,67	0,52	0,61
rappel	0,90	0,88	0,93	0,94

Tableau 3. Performances du système pour l'extraction thématique pour les domaines cinéma et restaurant pour leurs critères respectifs acteur et scénario , service et propreté

Le Tableau 3 met en évidence que l'approche est très performante pour les quatre critères étudiés : le FScore est compris entre 0,67 et 0,89.

5.2. Performances du système en extraction d'opinion sur chacun des critères

Nous proposons ici de montrer les performances du système pour la détection d'opinion. Nous reportons les résultats obtenus pour le domaine du cinéma. Les indicateurs classiques de validation utilisés sont le rappel, la précision et le FScore. Un extrait est classé positif, négatif ou neutre selon la valeur de la somme des scores des mots calculés par *SenticNet*.

Performances du système pour l'extraction d'opinion sur le domaine « cinéma »						
	critère *acteur*			critère *scenario*		
	positif	*neutre*	*négatif*	*positif*	*neutre*	*négatif*
FScore	0,67	0,46	0,25	0,70	0,41	0,21
précision	0,53	0,63	0,50	0,57	0,67	0,45
rappel	0,91	0,36	0,17	0,92	0,30	0,14

Tableau 4. Performances du système pour l'extraction d'opinion sur le domaine du cinéma pour les critères acteur et scénario

Le Tableau 4 regroupe les performances du système de détection d'opinion pour les critères *acteur* et *scenario*. Nous constatons que le système a des performances satisfaisantes pour la détection des opinions positives (FScore proche de 0,70). En revanche, on remarque une certaine faiblesse pour la détection des opinions négatives (FScore proche de 0,25). En effet, l'expression de jugements négatifs prend des formes syntaxiques complexes (« Je ne pense pas que cette comédie soit exceptionnelle ! ») qui nécessiterait l'utilisation d'une analyse syntaxique conjointement à *SenticNet*.

6. État de l'art

L'approche décrite dans cet article a pour objectif d'identifier des fragments de textes en relation avec une thématique donnée et d'extraire l'opinion qui lui est propre. Ainsi, nos travaux ont de forts liens avec les tâches de segmentation de textes, d'extraction thématique et de détection d'opinion.

La segmentation de textes vise à identifier les ruptures thématiques dans un document afin de le découper en extraits homogènes. Ces derniers sont considérés comme des « morceaux de textes » comportant de forts liens sémantiques internes, tout en étant détachés des extraits adjacents. Dans notre approche plus globale décrite en introduction, la recherche de structure thématique (McDonald *et al.*, 2002 ; Misra *et al.*, 2011) représente une étape préalable cruciale. De manière similaire à de nombreux travaux, nous nous sommes appuyés sur des méthodes statistiques. Par exemple, TextTilling (Hearst, 1997) étudie la distribution des termes suivant des critères. La répartition des termes se révèle en effet une information essentielle qui est souvent utilisée pour la segmentation (Reynar, 2000). D'autres méthodes, comme l'approche C99 (Choi, 2000), sont fondées sur le calcul de similarité entre les phrases afin de les rassembler et de détecter les ruptures thématiques. Notons que les approches de segmentation présentent, selon nous, une faiblesse majeure ; elles ne sont pas capables d'identifier la thématique précise d'un extrait.

Pour résoudre cette problématique d'étiquetage, des techniques issues du résumé de textes permettent, quant à elles, d'identifier les parties d'un document en fonction de la thématique dominante (Chuang *et al.*, 2000). D'autres méthodes visent à identifier des extraits relatifs au titre (Kupiec *et al.*, 1995). La majorité des techniques de résumé automatique s'appuie sur des méthodes d'apprentissage supervisé qui nécessitent une intervention humaine pour la constitution d'un corpus d'entraînement conséquent.

Notre approche utilise des informations statistiques comme la majorité des méthodes de segmentation et est capable d'identifier, dans un cadre non

supervisé, des segments de textes en rapport avec une sous-thématique. La seule intervention humaine réside dans la définition du critère par un nombre restreint de mots.

En outre, ce travail s'intéresse à la détection d'opinion, de sentiment, dans des données textuelles. Les premiers travaux remontent à la fin des années 1990 (Argamon *et al.*, 1998 ; Kessler *et al.*, 1997 ; Spertus, 1997), mais c'est seulement dans le début des années 2000 que cette problématique est introduite dans le traitement de l'information (Chaovalit *et al.* 2005 ; Dimitrova *et al.*, 2002 ; Durbin *et al.*, 2003). Jusqu'au début des années 2000, les deux principales approches populaires à la détection d'opinion étaient basées sur des techniques d'apprentissage machine et sur des techniques d'analyse sémantique. Par la suite, des techniques, plus fines, de traitement du langage naturel ont été largement utilisées, notamment dans la détection d'opinion dans des documents. L'opinion-mining est à ce jour une discipline au carrefour du traitement du langage naturel et de la recherche d'informations, et comme tel, elle partage un certain nombre de techniques propres à chacune des deux disciplines, comme l'extraction d'informations et le text-mining.

On distingue deux types de classification d'opinion :

– la classification binaire des documents en apposant une étiquette (positif/négatif) sur chacun d'eux.

– la classification multiclasse qui offre plusieurs degrés d'opinion, ou échelle de valeur (fortement positif, positif, neutre, négatif, fortement négatif).

La plupart des travaux sur l'analyse d'opinions a surtout mis l'accent sur une classification binaire des documents mais il est souvent utile d'avoir plus qu'une information binaire, surtout quand l'objectif final est un système de recommandation ou de confrontation d'opinions (Xu *et al.*, 2011).

À ce jour ces nouvelles techniques d'opinion-mining sont largement utilisées dans des applications extrêmement variées comme la comparaison de produits dans le e-commerce où l'objectif est la recommandation de produits (Bai, 2011 ; Pang *et al.*, 2002 ; He *et al.,* 2010). Dans ce contexte, on se rend compte que chaque produit dispose la plupart du temps de plusieurs fonctionnalités, dont seulement une partie d'entre elles sont susceptibles d'intéresser le consommateur (Morinaga *et al.*, 2002 ; Taboada *et al.,* 2006). Nous nous intéressons dans ce cas à détecter une opinion personnalisée, en fonction des besoins de l'utilisateur. Cette personnalisation permet une détection d'opinion raffinée, très appréciée dans les systèmes d'analyse et de comparaison des opinions de consommateurs sur des produits concurrents, c'est ce que propose (Liu *et al.*, 2005) avec un système prototype appelé Opinion Observer. Il se décompose en deux étapes, la première est l'identification des caractéristiques

du produit fondé sur des techniques de traitement automatique du langage et l'extraction de motifs. Ces caractéristiques constituent la base des critères de comparaison. La seconde étape consiste à émettre une opinion binaire (positif/négatif) pour chacune des caractéristiques (ou critère) précédemment identifiées. Dans Jin *et al.*, (2009), les auteurs proposent le système OpinionMiner dont l'objectif est d'extraire des entités de produits, i.e. des fonctions spécifiques liées à un produit et les opinions qui y sont associées. L'objectif de cette méthode est proche de celle présentée dans cet article.

Une autre application à l'opinion-mining est le résumé d'article d'opinion (Ku *et al.*, 2005 ; Beineke *et al.*, 2004) qui a pour objectifs d'analyser les tendances des utilisateurs (Bai, 2011), de détecter des produits phares, d'effectuer des retours clients... Ces techniques de résumé visent à identifier la polarité (positif/négatif), le degré et les évènements corrélés d'un document. (Hu *et al.*, 2004) proposent une technique d'analyse de commentaires d'internautes relatif à un produit particulier, la technique se résume en trois points : 1) identifier les caractéristiques du produit où les clients ont exprimé leur opinion, 2) pour chaque caractéristique, identifier les phrases ou avis émis (positifs ou négatifs), et 3) produire un résumé en utilisant les informations découvertes.

(Cardie *et al.*, 2003 ; Clarke *et al.*, 2003 ; Gopal *et al.*, 2011) proposent une détection d'opinion argumentée (Opinion reason mining) en appuyant leur jugement par l'extraction des phrases (ou extrait) qui leur a permis d'émettre un tel jugement.

Dans un contexte politique, Thomas *et al.*, (2006) tentent de déterminer à partir de la transcription des débats du Congrès des États-Unis si les discours représentent un appui ou une opposition à la législation proposée. Mullen *et al.*, (2006) décrivent une méthode statistique d'analyse de sentiment politique sur des discussions de groupes politiques pour déterminer s'ils sont en opposition avec le message original.

La tâche de classification de document par l'opinion qui s'en dégage implique une construction manuelle ou semi-manuelle d'un lexique de mots d'opinion, construit par des techniques de classification (Hatzivassiloglou *et al.*, 1997 ; Lin, 1998 ; Pereira *et al.*, 1993). La classification de mots ou phrases par leur orientation sémantique est positive ou négative, ou parfois associée à une intensité de jugement. Elle est généralement obtenue en utilisant un ensemble présélectionné de mots germes, ou en utilisant des heuristiques linguistiques (par exemple, (Lin, 1998) et (Pereira et *al.*, 1993)). Certaines études ont montré que restreindre la classification aux seuls adjectifs améliore les performances (Andreevskaia *et al.*, 2006 ; Turney *et al.*, 2002 ; Wiebe *et al.*, 2005). Outre le fait

que cela se vérifie pour certains cas particuliers, la majorité des méthodes existantes admet que cette restriction à la seule considération des adjectifs comme mots porteurs d'opinion est insuffisante, et qu'il est nécessaire de considérer aussi les adverbes, quelques noms et les verbes car ils sont porteurs d'une orientation sémantique (Andreevskaia *et al.*, 2006 ; Esuli *et al.*, 2005). On peut distinguer deux méthodes d'annotation automatique de mots porteurs d'opinion : les approches basées sur un corpus d'apprentissage et les approches basées sur l'utilisation d'un dictionnaire. Les méthodes basées sur un corpus annoté s'appuient généralement sur une analyse syntaxique et de co-occurrence des mots (Hatzivassiloglou *et al.*, 2000 ; Turney *et al.*, 2002 ; Hong *et al.*, 2003). Les autres utilisent un dictionnaire WordNet (http://wordnet.princeton.edu/) et s'appuient sur les informations fournies par l'outil pour obtenir l'orientation sémantique d'un mot (Xu *et al.*, 2011 ; Kim *et al.,* 2004), ou mesurent la similarité entre des mots candidats et les mots porteurs d'opinion comme bad et good par exemple (Kamps *et al.*, 2004).

Même si les problématiques abordées dans les articles précédemment énoncés peuvent sembler proches de la nôtre (i.e. les entités peuvent correspondre par exemple aux caractéristiques, aux fonctions ou aux composants d'un appareil photo), ces approches supposent une étape d'apprentissage nécessitant qu'un expert spécifie, dans un grand volume de documents, toutes les phrases correspondant aux entités. Dans notre cas, cette approche est automatique. Nous pensons que cette intervention experte est une contrainte majeure qui nuit à une diffusion plus large des techniques d'opinion-mining. En ce qui concerne la détection d'opinion, les techniques utilisent pour la plupart des approches supervisées (Yi *et al.*, 2003 ; Oelke *et al.*, 2009) : les mots d'opinion sont définis soit par des dictionnaires: WordNet, General Inquirer, Dictionary of Affect of Language (DAL), soit manuellement. D'autres techniques non supervisées introduisent la notion d'apprentissage des termes d'opinion à partir de mots germes pour constituer leur propre dictionnaire d'opinion (Harb *et al.*, 2008 ; Oelke *et al.*, 2009) de façon automatique. Notre approche, s'inspirant des techniques utilisées pour la détection d'opinion précédemment citées, les transposent pour l'extraction thématique de textes et l'extraction d'opinion.

7. Conclusion

Dans cet article nous avons proposé une nouvelle approche permettant de caractériser de manière automatique des critères pour un domaine donné et d'extraire l'opinion qui leur est relative. Nous avons montré que les mots caractéristiques d'un critère ne sont pas suffisants pour réaliser une extraction thématique efficace, et qu'il est indispensable de considérer les mots du

domaine non caractéristiques du critère (i.e. l'anti-classe). Nous avons également proposé une méthode permettant la construction automatique de corpus nécessaires à la construction de lexiques de mots représentatifs de critères, utilisés pour l'extraction automatique de segments de textes. Par ailleurs, à partir de ces segments, nous proposons d'en extraire automatiquement les opinions relatives à un critère choisi. Enfin, nous avons obtenu des performances remarquables pour l'extraction thématique ainsi que pour l'extraction d'opinion relative à un critère.

Les perspectives associées à ce travail sont nombreuses. Tout d'abord, nous avons choisi d'utiliser le lexique dans un contexte d'extraction thématique pour des données textuelles, ceci pourrait aussi être la première étape de la construction automatique d'ontologies, ou être utilisé dans un contexte de classification de documents. Par ailleurs, nous souhaitons améliorer notre approche d'extraction d'opinion spécifique à des critères. Dans de précédents travaux (Harb *et al.*, 2008), nous avions mis en évidence, qu'en fonction du domaine les opinions s'expriment par des termes spécifiques. L'étude de ces vocabulaires spécifiques permettrait, sans aucun doute, d'affiner nos résultats en matière d'analyse d'opinion.

Bibliographie

Andreevskaia A., Bergler S., (2006). « Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses », *Proceedings EACL-06,* Trento, Italy.

Argamon S., Koppel M., Avneri Galit, (1998). "Routing Documents According to Style", *Proceedings of First International Workshop on Innovative Information Systems*.

Bai X. (2011). "Predicting consumer sentiments from online text", *Decision Support Systems*, 50 (4), p. 732-742.

Beineke P., Hastie T., Manning C., Vaithyanathan S., (2004). "Exploring Sentiment Summarization", *Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications* AAAI Press, p. 1-4.

Cambria E., Speer R., Havasi C., Hussen A., (2010). "SenticNet: A Publicly Available Semantic Resource for Opinion Mining", *Artificial Intelligence*, p. 14-18.

Cardie C., Wiebe J., Wilson T., Litman D. (2003). "Combining Low-Level and Summary Representations of Opinions for Multi-Perspective Question Answering", *Proceedings of the AAAI Spring Symposium on New Directions in Question Answering*, p. 20-27.

Castro-Schez J.J., Miguel R., Vallejo D., Lopez-Lopez L.M. (2011). "A highly adaptive recommender system based on fuzzy logic for B2C e-commerce portals", *Expert Systems with Applications*. 38 (3), p. 2441-2454.

Chaovalit P., Zhou L., (2005). "Movie review mining: a comparison between supervised and unsupervised classification approaches", *Hawaii International Conference on System Sciences*, Los Alamitos, CA, USA: IEEE Computer Society 4, p. 112c.

Choi F. (2000). "Advances in domain independent linear text segmentation", *ACL'00*. 23, p. 26-33.

Chuang W.T., Yang J., (2000). "Extracting Sentence Segments for Text Summarization: A Machine Learning Approach", *Proceedings of the 23 th ACM SIGIR,*. p. 152-159.

Clarke C.L.A., Terra E.L. (2003). "Passage retrieval vs. document retrieval for factoid question answering", *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, New York, NY, USA: ACM, p. 427-428.

Dimitrova M., Finn A., Kushmerick N., Smyth B., (2002). "Web Genre Visualization" *Proc. conference on human factors in.*

Durbin S.D., Richter J.N., Warner D. (2003). "A system for affective rating of texts" *Proceedings of OTC-03, 3rd workshop on operational text classification*, Washington, USA.

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

Garcia I., Sebastia L., Onaindia E. (2011). « On the design of individual and group recommender systems for tourism", *Expert Systems with Applications*, 38 (6), p. 7683-7692.

Gopal R., Marsden J.R., Vanthienen J., (2011). "Information mining — Reflections on recent advancements and the road ahead in data, text, and media mining", *Decision Support Systems.*

Harb A., Plantié M., Dray G., Roche M., Trousset F., Poncelet P., (2008). "Web opinion mining: how to extract opinions from blogs?", *International Conference on Soft Computing as Transdisciplinary Science and Technology.*

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

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

He Y., Zhou D., (2010). "Self-training from labeled features for sentiment analysis", *Information Processing & Management.*

Hearst MA. (1997). "Text-tilling: segmenting text into multi-paragraph subtopic passages", *Computational Linguistics*, p. 59-66.

Hong Y., Hatzivassiloglou V., (2003). "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences", *Proceedings of EMNLP-03*, p. 129-136.

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

Jin W., Ho Hung H., Srihari R.K (2009). "OpinionMiner: A Novel Machine Learning System for Web Opinion Mining and Extraction", *IEEE Symposium on Visual Analytics Science and Technology*.

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

Kessler B., Numberg G., Schütze H., (1997). „Automatic detection of text genre" *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, p. 32-38.

Kim S.-M., Hovy E., (2004). "Determining the sentiment of opinions", *Proceedings of the 20th international conference on Computational Linguistics*. Stroudsburg, PA, USA, Association for Computational Linguistics.

Kleiber G. (1996). « Noms propres et noms communs : un problème de dénomination », *Meta*, Presses de l'Université de Montréal, p. 567-589.

Ku L.-W., Li L.-Y., Wu T.-H., Chen H.-H., (2005). "Major topic detection and its application to opinion summarization", *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*, p. 627-628.

Kupiec J., Pedersen J., Chen F. (1995). "A trainable document summarizer", *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM p. 68-73.

Lin D., (1998). "Automatic retrieval and clustering of similar words", *Proceedings of the 17th international conference on Computational linguistics* - Volume 2, Stroudsburg, PA, USA, Association for Computational Linguistics, p. 768-774.

Liu B., Hu M., Cheng J., (2005). "Opinion observer: analyzing and comparing opinions on the Web", *Proceedings of the 14th international conference on World Wide Web*, New York, NY, USA: ACM, p. 342-351.

McDonald D., Chen H. (2002). "Using sentence-selection heuristics to rank text segments in TXTRACTOR", *JCDL'02*, ACM Press, p. 28-35.

Misra H., Yvon F., Cappé O., Jose J. (2011). "Text segmentation: A topic modeling perspective", *Information Processing & Management*.

Morinaga S., Yamanishi K., Tateishi K., Fukushima T. (2002). "Mining product reputations on the web", *ACM SIGKDD*., p. 341-349.

Mullen T., Malouf R. (2006). "A preliminary investigation into sentiment analysis of informal political discourse", *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW*, p. 159-162.

Oelke D., Hao M.C., Rohrdantz C., Keim D.A., Dayal U., Haug L.-E. Janetzko H., (2009). "Visual Opinion Analysis of Customer Feedback Data", *KDD'09*.

Pang B., Lee L., (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of EMNLP*, p. 79-86.

Pereira F., Tishby N., Lee L. (1993). "Distributional clustering of English words", *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA, Association for Computational Linguistics, p. 183-190.

Reynar J. C. (2000). *Topic segmentation: Algorithms and applications,* Seattle.

Schmid H., (1994). "TreeTagger", *TC project at the Institute for Computational Linguistics of the University of Stuttgart*.

Spertus E. (1997). "Smokey: Automatic recognition of hostile messages", *Proc. of Innovative Applications of Artificial Intelligence (IAAI*), p. 1058-1065.

Taboada M., Gillies M.A., McFetridge P., (2006). "Sentiment Classification Techniques for Tracking Literary Reputation", *LREC Workshop: Towards Computational Models of Literary Analysis*, p. 36-43.

Thomas M., Pang Bo, Lee Lillian (2006). "Get out the vote: Determining support or opposition from Congressional floor-debate transcripts", *Proceedings of EMNLP*, p. 327-335.

Turney P., Littman M., (2002). "Unsupervised learning of semantic orientation from a hundred-billion-word corpus", *National Research Council of Canada.*

Wiebe J., Riloff E., (2005). "Creating Subjective and Objective Sentence Classifiers from Unannotated Texts", *Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Text Processing , CICLing.*

Xu K., Liao S.S., Li J., Song Y. (2011). "Mining comparative opinions from customer reviews for Competitive Intelligence", *Decision Support Systems*, 50 (4), p. 743-754.

Yi J. Nasukawa T., Bunescu R., Niblack W. (2003). "Sentiment Analyser: Extraction Sentiments about a Given Topic using Natural Language Processing Techniques", *IEEE Intl. Conf. on Data Mining (ICDM).*