Publié dans les
Actes 10e Journées Analyse statistique des Données Textuelles JADT 2010, 653-664, 2010
qui doivent être utilisés pour toute référence à ce travail

Classification automatique d'opinions dans la blogosphère

Jacques Savoy, Olena Zubaryeva

Institut d'informatique
Université de Neuchâtel – rue Emile Argand 11 - 2009 Neuchâtel - Suisse

Abstract

This paper describes the problem of classifying opinion from blogs. After retrieving relevant sentences, the search system must categorize them as opinionated or factual. To achieve this objective, different representations and automatic categorization models could be used. As baseline system, we have used the *Naïve Bayes* approach to classify the retrieved sentences as opinionated or not. As a second model, we have used an SVM model (based on a *tf idf* representation) showing an increase in the overall performance. We suggest using a normalized score (Z score) for each term according to its presence or absence in opinionated sentences. Based on these Z-scores we can determine whether a given sentence belongs to opinionated or not-opinionated category. The proposed system has been evaluated using the NCTIR English test-collection. We show that the suggested classification method performs significantly better than other approaches. Using a specialized thesaurus, we can further improve the overall categorization performance.

Résumé

Cette communication traite de la classification automatique d'opinions dans la blogosphère. Sur la base d'une liste de phrases jugées pertinentes, le système doit déterminer si elles contiennent une opinion ou non. Afin d'atteindre ce but, plusieurs représentations et modèles de catégorisation peuvent être utilisés. Comme système de référence, nous avons retenu une classification basée sur le modèle *Naïve Bayes*. L'emploi d'une stratégie SVM (avec une représentation *tf idf*) permet d'accroître la performance du système. Le système que nous proposons détecte l'usage d'un vocabulaire spécifique à chaque catégorie en recourant à un score normalisé (score z). Ces valeurs nous permettent de déterminer si une phrase contient ou non une opinion. Le système proposé a été implémenté et évalué grâce à la collection test NCTIR écrite en langue anglaise. Cette évaluation indique que notre modèle apporte clairement la meilleure performance. Le recours à un thesaurus spécialisé permet d'accroître encore la performance de catégorisation.

Mots-clés : détection d'opinions; classification d'opinions ; évaluation de classifieurs ; RI dans la blogosphère.

1. Introduction

Au fil des années, le rôle des internautes s'est modifié. Ils ne se contentent plus d'être de simples consommateurs mais participent et collaborent à l'enrichissement des sites Internet à l'image de l'encyclopédie Wikipédia. Comme autres exemples, on peut citer la diffusion de son journal, rendre public son opinion personnelle, écrire son carnet de bord d'artiste ou partager ses émotions face aux évènements. Tous ces exemples se rencontrent dans les forums, les sites de discussion, les réseaux sociaux ou, plus généralement, dans les *blogs*.

Il s'avère illusoire de dresser un inventaire exhaustif des thèmes abordés par les *blogs* (ou la *blogosphère*) car ils couvrent toutes les activités et préoccupations humaines. Cependant, si chaque *blog* possède, originellement pour le moins, un caractère autobiographique, on peut attribuer à ces journaux électroniques les qualificatifs de "subjectif" et "d'opinion". Parfois centré exclusivement sur une personne (*e.g.*, un artiste, un homme public) ou un produit (*e.g.*, la voiture, le football), le *blog* possède très souvent un aspect interactif et participatif. Ainsi

chaque billet (*post*) publié peut faire l'objet de commentaires de la part des lecteurs, parfois de manière continue.

La *blogosphère* se caractérisant par un contenu plus subjectif, les moteurs commerciaux proposent une entrée distincte pour y dépister des informations (*e.g.*, blogsearch.google.com). Les internautes y trouveront des informations ("Comment réparer la porte du Miele 120C") ou des opinions sur des produits ("l'écran du dernier Nokia est d'une qualité exceptionnelle"), personnes ou évènements. Les entreprises et les acteurs du marketing ont compris qu'elles pouvaient en tirer parti pour influencer, apprendre, communiquer, informer, se montrer, vendre (Malaison, 2007). Les acteurs politiques ont également suivi cette tendance avec des différences notables entre les Etats-Unis et la France, voire entre politiciens français (Véronis *et al.*, 2007), (Greffet & Wojcik 2008).

Mais cet espace communautaire soulève également de nouveaux défis. Ainsi, les fautes d'orthographe et d'accord connaissent une plus grande fréquence et le recours à un langage SMS (*e.g.*, "sui arivé paC 11h") requiert des pré-traitements appropriés (Boiy & Moens, 2009). La présence de pourriels (*spams*), de mensonges, de propagande ou de manipulation d'opinion rend le contenu de la blogosphère plus difficile à filtrer (Abassi *et al.*, 2008).

Dans cet article nous désirons proposer un modèle de catégorisation d'opinions extraites de blogs. Après une présentation générale des travaux reliés (section 2), la troisième section décrit le corpus de référence utilisé et les mesures de performance adoptées. La quatrième section expose les divers modèles de catégorisation que nous avons implantés et leur évaluation. La dernière section résume les principales contributions de cette communication.

2. Dépistage et détection d'opinions

La nature plus subjective de la blogosphère par rapport à d'autres parties de la Toile a très vite attiré l'attention des chercheurs. Malgré les différences, par rapport aux pages web traditionnelles, les modèles de recherche d'information les plus performants pour la blogosphère s'avèrent les mêmes que ceux utilisés dans d'autres domaines. Lors des campagnes d'évaluation TREC 2006 (Ounis *et al.*, 2007) et TREC 2007 (MacDonald *et al.*, 2008) ainsi que sur la base d'études complémentaires (Fautch & Savoy, 2008), on peut toutefois indiquer que la suppression des suffixes flexionnels et dérivationnels (*e.g.*, proposée par l'algorithme de Porter) diminue significativement la performance moyenne. Pour la langue anglaise pour le moins, la suppression automatique des suffixes même limitée à la forme pluriel en '-s' (Harman, 1991) ne s'avère pas toujours être une stratégie adéquate. Par contre, favoriser les billets d'information possédant conjointement plusieurs mots en commun avec la requête ou ceux dont les mots recherchés apparaissent proches les uns des autres constituent des techniques améliorant significativement la performance. Enfin l'élargissement automatique de la requête sur la base de mots extraits de Wikipédia, de sites dédiés aux nouvelles d'agence ou, plus généralement, du web permet d'accroître la qualité de la réponse.

Les billets pertinents aux souhaits de l'usager étant extraits, nous devons encore détecter s'ils contiennent une opinion (subjective), des sentiments ou, au contraire, s'ils correspondent simplement à une description. Dans cette perspective, on peut distinguer essentiellement deux niveaux de granularité. Plusieurs travaux, comme notre étude, s'appuient uniquement sur la phrase comme élément atomique pour détecter des opinions. Comme alternative, la détection d'opinion s'effectue au niveau du document et une telle classification s'appuie sur les composantes internes que sont les paragraphes et phrases, par exemple, en calculant un score d'opinions comme la somme des scores des entités individuelles (Gerani *et al.*, 2009).

Afin de déterminer si une phrase contient ou non une opinion, les systèmes de classification proposés se distinguent entre les approches s'appuyant sur le lexique (Elusi & Sebastiani, 2006) ou celles basées sur un apprentissage automatique (Abassi *et al*., 2008). Dans le premier cas de figure, le système recherche des mots caractéristiques (ou des parties du discours) afin de permettre une classification. Ainsi, Levin propose de définir différentes catégories verbales (déclaration, caractérisation, conjecture, admiration, jugement, élocution, etc.) ainsi que leurs verbes et formes introductives caractéristiques. Par exemple, une émotion peut être annoncée par la forme "John *was surprised* when …" ou une clause explicative suit la préposition *because* (e.g., "The man walks on the moon *because*…"). Dans le cadre de la détection d'opinion et se basant sur ces principes, on peut citer, par exemple, les travaux de Bloom *et al*. (2007). Comme autre exemple, on peut également citer Harb *et al*. (2008) utilisant la liste {"good", "nice", "excellent", "positive", "fortunate", "correct", "superior"} comme adjectifs pouvant indiquer une opinion positive et l'ensemble {"bad", "nasty", "poor", "negative", "unfortunate", "wrong", "inferior"} pour une opinion négative.

S'appuyant sur des modèles d'apprentissage par machine, plusieurs auteurs (Abassi *et al*., 2008), (Boiy & Moens, 2009) suggèrent d'adopter une représentation des phrases sous la forme d'une séquence de termes. Ensuite, sur la base d'un corpus d'entraînement composé d'exemples et de contre-exemples, le système apprend et fixe la valeur des paramètres d'un modèle de catégorisation (Sebastiani, 2002). Dans ce cadre, le modèle *Naïve Bayes* (Mitchell, 1997), machines à vecteurs de support (Joachims, 2002) ou des modèles de langue (Pang & Lee, 2004) sont les approches les plus usitées sans que l'on sache clairement quelle méthode propose les meilleures performances (Seki *et al*., 2008).

Toutes ces solutions doivent répondre à des questions récurrentes à tout traitement automatique de la langue naturelle. Doit-on ignorer les mots-outils (déterminant, conjonctions, prépositions, pronoms, etc.) ? Comment doit-on pondérer la valeur des divers mots présents dans une phrase ? Doit-on recourir à un traitement morphologique ou un simple enracineur ? Comment les variantes morphologiques et morphosyntaxiques influencent-elles la performance ? La détection des groupes nominaux (*e.g*., "batterie de l'iPhone", "téléphone portable") permet-elle d'améliorer la qualité du système ? Les pratiques stylistiques peuvent substituer un nom par un synonyme ou par un pronom présentant ainsi un obstacle à l'établissement d'une statistique telle que les fréquences d'occurrence (ou *tf*) (Minel, 2002).

3. Les campagnes d'évaluation NTCIR-6 & NTCIR-7

3.1. Les collections de test

Afin de disposer d'un corpus d'évaluation pour la détection et la classification d'opinions, les organisateurs de la campagne NTCIR-6 (Seki *et al*., 2007) puis NTCIR-7 (Seki *et al*., 2008) ont sélectionné des articles extraits de journaux parus dans les années 1998 à 2001 comme, pour la langue anglaise, le *Mainichi Daily News*, *Korea Times*, *Xinhua News*, *Hong Kong Standard*, *Straits Times*. Il existe également une collection similaire de phrases écrites en langue japonaise et chinoise traditionnelle. Ces campagnes ont également pour vocation de favoriser l'application d'approches différentes sur le même jeu de données afin de mieux connaître leurs avantages et défaillances.

Pour définir les documents puis les phrases formant le référenciel, les articles retenus devaient répondre à des besoins d'information exprimés couvrant des sujets divers comme la politique ("President of Peru, Alberto Fujimori, scandal, bribe"), le sport ("Jennifer Capriati, tennis"), ou la société ("Divorce, Family Discord, Criticisms", "History Textbook, Controversies,

World War II"), touchant parfois des sujets plutôt nationaux ("Kim Dae Junm Kim Jung II, Inter-Korean Summit") ou, inversement, des thèmes possédant une couverture internationale ("Animal Cloning Techniques" ou "Nuclear power protests").

Les articles répondant à ces diverses interrogations étant connus, les organisateurs ont sélectionné entre cinq et environ vingt articles par requête. Sur cette base, ils ont ensuite procédé à une segmentation en phrases pour définir à ce niveau la présence ou non d'une opinion, puis à une classification de cette opinion comme positive, négative ou neutre.

Si l'on regroupe les deux corpus NTCIR-6 et NTCIR-7, nous disposons de 10 145 phrases dont 7 650 (ou 75,4 %) ne renferment pas d'opinion. Le reste (2 495 phrases) contient, aux yeux des assesseurs, une opinion. Cette dernière peut être positive (dans 480 cas ou 4,7 %), négative (1 064 cas, 10,5 %) ou neutre (951 phrases, 9,4 %). Cette dernière catégorie regroupe parfois une phrase présentant une opinion positive sur un aspect et négative sur un autre.

Le tableau 1 reprend quelques exemples. Au haut de ce tableau, on retrouve une phrase extraite du journal *Hong Kong Standard* (information donnée par l'étiquette <DOCID>). Ensuite on a indiqué par une valeur booléenne si la phrase contient une opinion (<OP> Y </OP>) ou non (<OP> N </OP>). Si la phrase renferme une opinion, la polarité de cette dernière est indiquée comme positive (étiquette <POL> POS </POL>), négative (<POL> NEG </POL>) ou neutre (<POL> NEU </POL>). Finalement la phrase apparaît entre la balise <TEXT>.

<DOCID> HK-199906230280191.0034.E </DOCID> <OP> N </OP>
<TEXT> ``The rains have been falling since Saturday and much of the town is under water now," police said.

<DOCID> EN-9910262A11LH315.0002.E </DOCID> <OP> Y </OP> <POL> NEG </POL>
<TEXT> Carl (not his real name) had been "on the low ebb," as he put it, over the past few weeks.

<DOCID> EN-9910262A11LH315.0006.E </DOCID> <OP> Y </OP> <POL> POS </POL>
<TEXT> Away from friends and family, Japan, he thought, would be the place to let him focus on putting the pieces back to together.

<DOCID> HK-199906230280191.0014.E </DOCID> <OP> Y </OP> <POL> NEU </POL>
<TEXT> ``For us, health and safety issues always come before business issues," Ivester said.

<DOCID> HK-199912200280034.0022.E </DOCID> <OP> Y </OP> <POL> NEU </POL>
<TEXT> George Amato of the Wildlife Conservation Society's Science Resource Centre and the Bronx Zoo in New York City said that cloning is one tool that can be used to save rare animals, but he's not keen on it for his own research.

***Tableau 1** : Exemples de phrases du corpus d'évaluation, la première étant sans opinion puis une phrase contenant une opinion négative, une positive et enfin deux neutres*

Noter une opinion comme négative dans l'exemple "This movie isn't very good" ou positive dans "The iPhone is a beautiful smartphone" ne soulève *a priori* pas de débat. Mais l'expression d'une opinion n'est pas toujours aussi nette et deux personnes peuvent avoir des jugements différents sur l'une ou l'autre phrase. Ainsi chaque phrase des corpus a été jugée par au moins trois annotateurs, sans forcément être toujours les mêmes personnes pour toutes les phrases. Selon la majorité des jugements, la phrase est jugée comme contenant ou non une opinion.

Cet ensemble (phrases, jugements) nous permet d'évaluer et d'analyser plusieurs stratégies de classification. Dans ce but, nous nous sommes concentrés sur la détection d'opinions sans considérer leur polarité (positive, négative ou neutre). En effet, en consultant ce corpus, il nous est apparu que l'affectation de la polarité était parfois sujette à discussion tandis que la

distinction entre “avec ou sans opinion” était plus claire. Avant de présenter les modèles de catégorisation, nous allons introduire les mesures de performance utilisées.

3.2. Mesures d’évaluation

Afin de mesurer la performance, les campagnes d’évaluation proposent de recourir à la précision, au rappel et, comme mesure combinée, la mesure F. Afin de calculer ces trois valeurs, nous établissons une table de contingence (voir tableau 2) dans laquelle on distingue les décisions prises par l'ordinateur (les lignes) et les jugements corrects (les colonnes). On indique le nombre de décisions de type “vrai positif” (phrase contenant une opinion et détectée comme telle par le système), “vrai négatif” (phrase sans opinion et catégorisée comme telle par le système) ainsi que les erreurs de catégorisation soit les “faux positif” (phrase sans opinion mais signalée comme renfermant une opinion) et les “faux négatif”.

	Jugement humain	
Décision de la machine	Opinion	Sans opinion
Positif	vrai positif	faux positif
Négatif	faux négatif	vrai négatif

***Tableau 2** : Exemple d'une table de contingence de résultat*

Sur la base de cette table, nous pouvons calculer la précision selon la formule 1. Cette valeur varie entre 0 (aucune détection) et 1,0 (toutes les phrases signalées comme ayant une opinion en possède bien une). Le rappel s'évalue selon l’équation 2 et varie également entre 0 et 1,0 (ou 0 et 100 %). Cette deuxième mesure permet de connaître la capacité du système à extraire toutes les phrases ayant une opinion.

$$\text{Précision } \pi = \frac{\#vrai\ positif}{\#vrai\ positif + \#\ faux\ positif} \quad (1)$$

$$\text{Rappel } \rho = \frac{\#vrai\ positif}{\#vrai\ positif + \#\ faux\ négatif} \quad (2)$$

Comme le rappel et la précision varient en sens inverse l’un de l’autre, il s’avère souvent utile de disposer d’une seule mesure pouvant refléter la capacité du système à dire uniquement la vérité (précision) et toute la vérité (rappel). La mesure $F_{(\beta=1)}$ décrite dans l’équation 3 sera utilisée dans cette perspective. Dans le cas présent où $\beta = 1$, on accorde autant d'importance à la précision qu'au rappel.

$$F_{(\beta)} = \frac{(1+\beta^2)\cdot\pi\cdot\rho}{\beta^2\cdot\pi+\rho} \quad \text{et avec } F_{(1)} = \frac{2\cdot\pi\cdot\rho}{\pi+\rho} \quad (3)$$

Enfin, nous devons également considérer que le système doit disposer d'un ensemble de phrases pour déterminer les paramètres de son modèle d'apprentissage. Si nous utilisons toutes les phrases pour l'apprentissage et ce même ensemble pour l'évaluation, la mesure de performance obtenue (évaluation “rétrospective”) sera trop optimiste et plus ou moins biaisée selon les modèles d'apprentissage. Afin de mesurer l'efficacité d'un système nous avons adopté la méthode de la validation croisée sur la base de dix blocs. Dans ce cas, le k^e bloc est réservé à l'évaluation et les k-1 autres blocs pour l'apprentissage (Sebastiani, 2002). Dans ce cas, les phrases utilisées lors de l'apprentissage ne seront pas employées lors de l'évaluation.

4. Modèle de catégorisation automatique et évaluation

4.1. Pré-traitement et réduction de l'espace des caractéristiques

Afin de classifier automatiquement une phrase comme contenant ou non une opinion, nous devons choisir une représentation interne adéquate. Comme élément pertinent, plusieurs auteurs suggèrent de retenir les mots, sans tenir compte de leur ordre d'apparition (hypothèse dite du "sac de mots"). Toutefois, on peut faire l'hypothèse que des variantes morphologiques liées à la syntaxe ne modifient que peu la sémantique. Dans cette optique, on peut appliquer un enraciniseur léger (ou *S-stemmer*) (Harman, 1991) afin d'éliminer la flexion '-s' du pluriel en anglais. Comme alternative, nous pouvons également recourir à une analyse morphologique plus poussée donnant pour chaque mot son lemme (ou entrée dans le dictionnaire). Dans ce but, nous avons recouru au logiciel d'étiquetage syntaxique automatique de l'Université de Stanford (Toutanova *et al*., 2003). Ainsi, au mot "said" nous pouvons faire correspondre le lemme "(to) say", tandis qu'au vocable "cats" correspond, après enracinisation le terme "cat". Pour la langue anglaise et en recherche d'information, l'application d'un enracineur ou le recours à un traitement morphologique apporte une qualité statistiquement similaire (Fautsch & Savoy, 2009).

Notons que dans notre étude, le terme "mot" devient ambigu. Dans notre contexte, nous avons réservé la désignation "vocable" pour désigner les formes distinctes (après élimination du suffixe '-s' notant le pluriel) tandis que le terme "mot" indiquera une forme de surface apparaissant dans un texte. Ainsi, dans la phrase "the cats saw the big cat", on compte six mots mais seulement quatre vocables.

Quelques précisions doivent encore être apportées. Dans notre système, toute majuscule débutant un mot sera remplacée par la minuscule correspondante (e.g., "Jobs" → "jobs"). Par contre si le mot s'écrit uniquement avec des majuscules, celles-ci seront conservées (e.g., "US", "DOJ" ou "AIDS"). Nous espérons ainsi regrouper sous la même entrée des formes quelque peu différentes mais désignant la même entité. Par contre, comme d'autres langues naturelles, l'anglais connaît des variantes orthographiques comme, "color", "colour" ou "center", "centre" que notre système ne regroupera pas sous la même entrée.

	Phrases avec opinion			Phrases sans opinion		
	tf	vocable	*df*	*tf*	vocable	*df*
1	536	said	529	772	said	754
2	422	not	398	646	not	609
3	290	he	254	552	he	487
4	201	we	169	423	japan	386
5	175	I	152	394	two	383
6	166	US	143	386	US	359
7	166	government	161	371	government	353
8	158	should	151	368	korean	314
9	153	more	139	354	korea	318
10	141	japan	126	329	other	315
11	139	world	132	329	after	323
12	138	chinese	119	325	more	311
13	133	korea	120	315	south	297
14	127	economic	123	311	economic	292
15	116	other	111	306	countr	292

***Tableau 3** : Vocables les plus fréquents (après élimination de 41 formes très courantes) dans les phrases renfermant une opinion (2 495 phrases) et sans opinion (7 650 phrases)*

Finalement, nous n'avons pas procédé à un traitement plus approfondi de la morphologie afin, par exemple, d'éliminer les suffixes dérivationnels (*e.g.*, “Japan” et “Japanese”). De même, nous n'avons par tenu compte des synonymes pour les regrouper sous une seule entrée (via, par exemple, le thésaursus *WordNet* (Fellbaum, 1998)). Le tableau 3 indique les quinze vocables les plus fréquents dans l'ensemble des phrases avec et sans opinion. Dans ce tableau, nous avons indiqué la fréquence lexicale ou d'occurrence (colonne *tf*) ainsi que le nombre de phrases ayant au moins une occurrence du vocable (colonne *df*).

Sur l'ensemble des 10 145 phrases de notre corpus, on compte 233 351 mots pour 14 027 vocables différents (pour 15 259 avant l'application d'un enracineur léger) ou 12 416 lemmes (avec les signes de ponctuation). Nous avons également ignoré 41 vocables très fréquents et peu porteurs d'information (e.g., “the”, “is”, “of”, “and”, “which”). De plus, en éliminant les termes apparaissant trois fois ou moins, l'espace se réduit de 14 027 à 5 065 vocables, soit une réduction de 63,9 %. C'est sur ce vocabulaire que nos diverses stratégies de catégorisation vont s'appuyer. L'élimination des vocables peu fréquents correspond d'abord à un souci de réduire sensiblement l'espace de représentation à nos classifieurs. On peut également mentionner que cet élagage élimine de nombreuses fautes d'orthographe présentes sous la forme de mots ayant une fréquence unitaire (*hapax*).

4.2. L’approche Naïve Bayes

Afin d'évaluer divers modèles de catégorisation, nous avons adopté comme première solution l'approche *Naïve Bayes* (Mitchell, 1997). Dans ce cas, le système de catégorisation choisira parmi les deux hypothèses possibles (h_0 = “sans opinion”, h_1 = “avec opinion”) celle qui retournera la valeur maximale selon la formule 4. Dans cette dernière, *t* indique le nombre de vocables inclus dans la phrase courante et t_j les différents termes apparaissant dans la phrase.

$$\text{Arg}\max_{h_i} \text{Prob}[h_i] \cdot \prod_{j=1}^{t} \text{Prob}[t_j \mid h_i] \qquad (4)$$

Les probabilités sous-jacentes doivent encore être estimées. Pour les probabilités *a priori* $\text{Prob}[h_i]$, cette estimation se base sur le rapport entre le nombre de phrases sans opinion (7 650), respectivement avec (2 495), et le nombre total de phrases dans le corpus (10 145). Pour les probabilités liées aux divers vocables, nous regroupons toutes les phrases appartenant à une catégorie (ensemble noté Text_{hi}). Sur la base de cet ensemble de taille n_{hi}, on estime les probabilités selon la formule 5. Celle-ci correspond au rapport entre la fréquence lexicale dans l'ensemble Text_{hi} (notée tf_{hi}) et la taille de l'ensemble correspondant.

$$\text{Prob}[t_j \mid h_i] = tf_{hi} \Big/ n_{hi} \qquad (5)$$

Cette estimation correspond au maximum de vraisemblance qui conduit à surestimer les probabilités des vocables présents dans le corpus au détriment des vocables absents. Dans ce dernier cas, la valeur tf_{hi} équivaut à 0, donnant une probabilité nulle d'occurrence. Or, il est reconnu que la distribution des mots suit une distribution de type LNRE (*Large Number of Rare Events* (Baayen, 2001)). Comme correction, un lissage simple consiste à ajouter une unité au numérateur de notre estimation et, en complément, d'ajouter au dénominateur la taille du vocabulaire retenu. Cette formulation se généralise (loi de Lidstone) en lissant toute probabilité par la formule $p = (tf_{hi}+\lambda) / (n_{hi}+\lambda \cdot |V|)$, avec λ un paramètre de lissage (fixé à 0,3) et $|V|$ indiquant la taille de notre vocabulaire (*e.g.*, 2 095 vocables si $h_i = 0$ et 4 137 si $h_i = 1$).

4.3. Séparateurs à vaste marge (SVM)

Au lieu de limiter la représentation des phrases par la présence ou l'absence de vocables, nous pourrions les pondérer pour refléter leur importance relative. Une telle approche se retrouve dans le modèle vectoriel en recherche d'information avec la pondération classique *tf idf* (Boughanem & Savoy, 2008). La composante *tf* indique la fréquence d'occurrence d'un terme dans la phrase. La valeur *idf* (= log(*df* / *n*)) correspond au logarithme de l'inverse de la fréquence documentaire (notée *df*). Cette valeur indique le nombre de phrases dans lesquelles ce terme apparaît. Finalement *n* désigne le nombre de phrases dans le corpus.

Comme alternative, nous pouvons normaliser les deux composantes afin qu'elles donnent des valeurs comprises entre 0 et 1. Pour la partie *tf*, nous avons implémenté la pondération *atf* = 0,5 + 0,5 · (*tf* / max *tf*). Le max *tf* représente la fréquence d'occurrence maximale dans la phrase considérée. Pour la partie *idf*, nous pouvons simplement diviser la valeur *idf* par log(*n*), normalisation que nous noterons *nidf*.

Disposant d'une représentation sous forme vectorielle, nous avons utilisé le système SVM^{light} (*Support Vector Machine*)[1] proposant un modèle d'apprentissage basé sur des séparateurs à vaste marge (Joachims, 2002). Dans ce cas le système détermine l'hyperplan séparant au mieux et de manière linéaire la représentation des phrases possédant une opinion de celles qui n'en n'ont pas.

4.4. Le score Z

Comme alternative à ces représentations, nous suggérons de pondérer les vocables en fonction de leur degré d'appartenance au vocabulaire spécifique (Muller, 1992) des phrases avec ou sans opinion. Afin de mesurer cette spécificité, on peut subdiviser notre corpus en deux, soit entre les phrases avec une opinion (ensemble noté O) et l'ensemble des phrases sans opinion (noté S). Pour un terme ω donné, on compte le nombre d'occurrences dans l'ensemble O (valeur notée *a* dans le tableau 4) et sa fréquence dans l'ensemble S (valeur notée *b*). Dans l'ensemble du corpus NTCIR (NTCIR = O ∪ S), nous aurons donc *a*+*b* occurrences de ce vocable. De manière similaire, la taille de l'ensemble O s'élève à *a*+*c* tandis que le volume du corpus NTCIR sera *n* = *a*+*b*+*c*+*d*.

	O	S	NTCIR
ω	*a*	*b*	*a*+*b*
Autres que ω	*c*	*d*	*c*+*d*
	a+*c*	*b*+*d*	*n* = *a*+*b*+*c*+*d*

***Tableau 4** : Exemple d'une table de contingence pour le vocable ω*

A l'aide du tableau 4 et pour le vocable ω, nous pouvons estimer sa probabilité d'occurrence dans l'ensemble NTCIR par Prob[ω] = (*a*+*b*) / *n* ou celle d'avoir un vocable provenant de l'ensemble O par Prob[O] = (*a*+*c*) / *n*.

Pour définir le pouvoir discriminant d'un vocable, nous faisons l'hypothèse que sa distribution suit une loi binomiale de paramètre Prob[ω] et *n'*. Prob[ω] désigne, lorsque nous tirons un mot du corpus, la probabilité de tirer le vocable ω. Cette probabilité s'estime par (*a*+*b*) / *n*. Si

[1] Disponible gratuitement à l'adresse http://svmlight.joachims.org/

l'on répète $n' = a+c$ fois ce tirage aléatoire, on peut estimer le nombre de vocable ω tiré par Prob[ω]·n'. Ce nombre correspond au nombre attendu et, dans le tableau 4, on retrouve le nombre réellement observé noté a. Une différence importante entre la valeur a et le produit Prob[ω]·n' indique que la distribution s'écarte du modèle de la binomiale. De manière précise, nous calculons un score Z pour chaque vocable ω selon l'équation 6 dans laquelle Prob[ω]·n' représente la moyenne et n'·Prob[ω]·(1-Prob[ω]) la variance de la binomiale.

$$score\ Z\ (\omega) = \frac{a - n' \cdot \mathrm{Prob}[\omega]}{\sqrt{n' \cdot \mathrm{Prob}[\omega] \cdot (1 - \mathrm{Prob}[\omega])}} \quad (6)$$

Comme règle de décision, on peut considérer que les vocables présentant un score Z supérieur à un seuil δ donné correspondent à un sur-emploi et ceux ayant un score inférieur à -δ à un sous-emploi. Pour l'ensemble des phrases ayant une opinion, les vocables possédant les plus fortes valeurs Z sont “should” (score Z = 10,12), “we” (7,61), “must” (6,83), “our” (6,06) ou “believe” (6,0). Inversement, dans les sous-emplois ou les vocables sur-employés dans les phrases sans opinion, on retrouve “last” (score Z = 3,63), “after” (3,38), “year” (3,29) ou “million” (3,27).

Afin de déterminer si une phrase contient une opinion, on récupère le score Z de chaque vocable de la phrase. Comme règle d'agrégation, nous calculons la somme des scores Z supérieurs à 1 (notée *sumPos*) et la somme des scores inférieurs à 1 (notée *sumNeg*). Si la *sumPos* > | *sumNeg* |, la phrase est placée dans la catégorie avec opinion (O), sinon dans celle des phrases sans opinion (S).

4.5. SentiWordNet

En se limitant à la langue anglaise, nous pouvons recourir au thésaurus *SentiWordNet*[2] (Esuli & Sebastiani, 2006) qui fournit, pour chaque lemme, trois valeurs numériques indiquant son degré de polarité positive, négative et objective. Par exemple, pour le terme “good”, ce thésaurus spécialisé redonne les valeurs de polarité positive = 0,875, négative = 0 et objective = 0,125 (la somme de ces trois valeurs numériques redonnant toujours l'unité).

Dans notre cas, nous nous limitons à déterminer si la phrase contient une opinion ou non. Nous pouvons réduire les indications de *SentiWordNet* à deux valeurs numériques. La présence d'une opinion sera indiquée par la somme des valeurs de polarité positive et négative tandis que l'absence d'opinion correspondra à la dernière valeur. Notons toutefois qu'un nombre important d'entrées possède une valeur d'objectivé unitaire (*e.g.*, “motor”, “Japan” ou “food”). De plus, certains termes ne sont pas inclus dans ce thésaurus (*e.g.*, “we” ou “Asian”) et seront donc ignorés par notre système de classification. En fait, la couverture de ce thésaurus (environ 1 000 synsets) demeure modeste par rapport à celle de *WordNet* (Fellbaum, 1998) et ses 115 000 synsets.

Afin de déterminer si une phrase contient une opinion, nous sommons, pour tous les vocables de la phrase, les scores de polarité et ceux d'objectivité. Comme de nombreux vocables inclus dans *SentiWordNet* retournent simplement une valeur d'objectivité unitaire, la somme des scores d'objectivité sera presque toujours supérieure à la somme des scores de polarité. Comme heuristique, nous divisons la somme des scores d'objectivité par le nombre de mots de la phrase. Si la somme des scores de polarité est supérieure à la moyenne des scores

[2] Disponible gratuitement à l'adresse http://SentiWordNet.isti.cnr.it/

d'objectivité, la phrase est étiquetée comme ayant une opinion, sinon nous estimons qu'elle appartient à la classe des phrases sans opinion.

4.6. Evaluation

Sur la base du corpus NTCIR (10 145 phrases dont 7 650 sans opinion et 2 495 avec), nous avons évalué les diverses stratégies de catégorisation sur la base d'une validation croisée (10 blocs). Les valeurs moyennes de précision, rappel et $F_{(1)}$ sont indiquées dans le tableau 5.

	Précision	Rappel	$F_{(1)}$
Naïve Bayes	18,89 %	67,45 %	29,52 %
SVM, vocables (*tf idf*)	33,63 %	65,76 %	44,37 %
SVM, vocables (a*tf nidf*)	34,99 %	64,97 %	45,33 %
SVM, lemmes (*tf idf*)	32,42 %	66,80 %	43,33 %
Score Z	**44,23 %**	**82,72 %**	**56,30 %**
SentiWordNet	82,34 %	53,25 %	64,68 %

***Tableau 5** : Evaluation de diverses stratégies de catégorisation (validation croisée, 10 blocs)*

Si l'on considère uniquement les approches ne nécessitant pas de connaissances additionnelles (soit le thésaurus spécialisé *SentiWordNet*), notre modèle basé sur le score Z propose la meilleure évaluation globale. Elle se détache si l'on considère la précision ou le rappel. Au niveau du de la mesure combinée $F_{(1)}$, le score Z surpasse l'approche SVM basée sur les vocables et une pondération *tf idf* (56,30 % vs. 44,37 %, différence relative de 26,9 %).

On constate que les variations des pondérations, dans le modèle SVM, ne modifient pas sensiblement la performance. Dans ce modèle SVM, l'emploi de fonctions noyaux polynomiales transformant l'espace de représentation des vocables (pondération *atf nidf*) améliore quelque peu la performance (45,33 % vs. 44,37 %, différence relative de 2,16 %). Finalement, pour ce modèle les différences entre une évaluation rétrospective (même ensemble pour l'entraînement et l'évaluation) et la validation croisée s'avèrent plus élevées. Par exemple, avec une pondération *tf idf* et une représentation par des vocables, la mesure $F_{(1)}$ passe de 44,37 % à 64,7 % (évaluation rétrospective), soit une augmentation relative de 45,9 %. On peut expliquer une telle différence par le fait que cette approche s'appuie sur un sous-ensemble de phrases afin de déterminer la catégorie. Lorsque la même phrase appartient à la fois au corpus d'entraînement et à celui d'évaluation, la décision s'avère plus simple pour ce modèle.

L'approche *Naïve Bayes* possède l'inconvénient de ne pas discriminer les vocables selon leur présence plus ou moins importante dans les phrases avec opinion ou sans opinion. En fait, l'estimation indiquée dans la formule 5 tient compte uniquement de la fréquence d'occurrence. Ainsi les vocables possédant les plus fortes probabilités sont “said” (prob. $1{,}68 \cdot 10^{-3}$), “not” ($1{,}32 \cdot 10^{-3}$) ou “he” ($0{,}91 \cdot 10^{-3}$) et ceci que ce soit dans l'ensemble des phrases avec ou sans opinion. Comme l'indique le tableau 3, ces termes sont fréquents dans les deux ensembles, et ceci indépendamment de la catégorie avec ou sans opinion.

Finalement, la performance du modèle basé sur le *SentiWordNet* démontre l'intérêt d'outils de traitement de la langue par rapport à des approches basées sur l'apprentissage automatique. Par contre, on doit reconnaître que cet outil n'est disponible que pour la langue anglaise. De plus, il s'avère bien adapté à notre besoin de distinguer entre phrases avec et sans opinion. En effet, la polarité de l'opinion n'est pas prise en compte. Ainsi, le système ne devra pas distinguer la polarité des opinions dans des séquences comme “beautiful” ou “not very

beautiful", penchant, dans les deux cas, pour l'expression d'une opinion. Distinguer entre une polarité positive ou négative peut tenir à un élément linguistique tel qu'un "not" et la performance de *SentiWordNet* ne serait pas, dans un tel cas de figure, forcément la meilleure.

5. Conclusion

La détection automatique d'opinion dispose de nombreuses applications, en particulier dans le suivi d'une clientèle prête à décrire ses expériences sur la Toile plutôt que de remplir de longs et fastidieux questionnaires. On notera également que l'acquisition de ces informations peut se faire très rapidement et que les opinions sont peu sujettes à être influencées de l'extérieur.

Dans cet article, nous avons proposé plusieurs représentations possibles d'une phrase, soit comme un ensemble de vocables ou de lemmes, avec ou sans pondération. Comme modèle de catégorisation, nous avons évalué l'approche Naïve Bayes ou le SVM, deux représentants emblématiques du paradigme de l'apprentissage automatique par machine. Notre modèle, basé sur le score Z, repose sur une différenciation entre les distributions lexicales dans les phrases avec ou sans opinion. Ce modèle permet d'apporter une qualité de réponse supérieure aux autres approches. Toutefois, l'usage d'un thésaurus spécialisé (*SentiWordNet*) permet d'accroître la performance générale dans la détection des opinions.

Comme perspective, nous pouvons combiner ces diverses approches afin d'améliorer la précision et le rappel. Enfin, nous pourrions tenter de classer les phrases renfermant une opinion en trois catégories distinctes, à savoir les phrases qui possèdent une opinion positive, négative ou mixte.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside n^0 200021-124389).

Références

Abassi A., Chen H. and Salem A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM-Transactions on Information Systems*, 26(3).

Baayen H.R. (2001). *Word Frequency Distributions.* Kluwer Academic Press, Dordrecht.

Bloom K., Stein S. and Argamon S. (2007). Appraisal extraction for news opinion analysis at NTCIR-6. *Proceedings NTCIR-6 (NII Test Collection for IR Systems)*, NII, Tokyo, pp. 279-289.

Boiy E. and Moens M.-F. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. *Information Retrieval*, 12(5), pp. 526-558.

Boughanem M. and Savoy J. (2008). *Recherche d'information. Etat des lieux et perspectives*. Hermès, Paris.

Bloom K., Stein S., Argamon, S. (2007). Appraisal extraction for news opinion analysis at NTCIR-6. *Proceedings NTCIR-6 (NII Test Collection for IR Systems)*, NII, Tokyo, pp. 279-289.

Esuli A. and Sebastiani F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings LREC-06 (Language Resources and Evaluation)*, pp. 417-422.

Fautsch C. and Savoy J. (2008). Stratégies de recherche dans la blogosphère. *Document numérique*, 11(1-2), pp. 109-132.

Fautsch C. and Savoy J. (2009). Algorithmic stemmers or morphological analysis: An evaluation. *Journal of the American Society for Information Sciences & Technology*, 60(8), pp. 1616-1624.

Fellbaum C. (1998). *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge (MA).

Gerani S., Carman M. J. and Crestani F. (2009). Investigating learning approaches for blog post opinion retrieval. *Proceedings ECIR 2009 (European Conference on IR Research)*, pp. 313-324.

Greffet F. and Wojcik S. (2008). *Parler politique en ligne.* Réseaux, Communication, Technologie, Société, Hermès, Paris.

Harb A., Plantié M., Roche M., Dray G., Trousset F. and Poncelet P. (2008). Détection d'opinion. *Document numérique*, 11(1-2), pp. 37-61.

Harman D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42, pp. 7-15.

Joachims T. (2002). *Learning to Classify Text Using Support Vector Machines. Methods, Theory and Algorithms.* Kluwer, London.

Macdonald C., Ounis I. and Soboroff I. (2008). Overview of the TREC-2007 blog track. *Proceedings TREC-2007 (Text REtrieval Conference)*, NIST Publication #500-274, pp. 1-13.

Malaison C. (2007). *Pourquoi bloguer dans un contexte d'affaires ?* Editions IQ, Montréal.

Minel J.-L. (2002). *Filtrage sémantique.* Hermès, Paris.

Mitchell T.M. (1997). *Machine Learning.* McGraw-Hill, New York.

Ounis I., de Rijke M., Macdonald C., Mishne G. and Soboroff I. (2007). Overview of the TREC-2006 blog track. In *Proceedings TREC-2006 (Text REtrieval Conference)*, NIST Publication #500-272, pp. 17-32.

Pang B. and Lee L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of ACL-2004 (Association for Computational Linguistics)*, Barcelona, pp. 271-278.

Sebastiani F. (2002). Machine learning in automatic text categorization. *ACM Computing Survey*, 14(1), pp. 1-27.

Seki Y., Evans D.K., Ku L-W., Chen H.-H. and Noriko K. (2007). Overview of opinion analysis pilot task at NTCIR-6. *Proceedings of NTCIR-6 (NII Test Collection for IR Systems)*, NII, Tokyo, pp. 265-278.

Seki Y., Evans D.K., Ku L-W., Sun L., Chen H.-H. and Noriko K. (2008). Overview of multilingual opinion analysis task at NTCIR-7. *Proceedings of NTCIR-7 (NII Test Collection for IR Systems)*, NII, Tokyo, pp. 185-203.

Toutanova K., Klein D., Manning C. and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclid dependency network. *Proceedings of HLT-NAACL 2003 (North American Chapter of the Association for Computational Linguistics – Human Language Technologies)*, pp. 252-259.

Véronis E., Véronis J. and Voisin N. (2007). *Les politiques mis au net.* Max Milo Editions, Paris.